Abstract

From Motor Learning to Social Learning:
A Study of Development on a Humanoid Robot

In this thesis, we describe how a humanoid robot designed to match the kinematics of a one-year old infant can learn to reach to visual targets, point toward visual targets, and share attention with a human. These three skills span the domains of motor learning and social learning. Instead of developing each of these skills independently, the robot naturally progresses from reaching to pointing and then from pointing to joint attention by taking advantage of social conventions and the assistance of a human.

We first present a biologically plausible model for learning to reach to visual targets. This model is then extended to allow the robot to point toward distant objects that are outside of its reach without requiring additional learning. We demonstrate that the development of joint attention can be greatly facilitated by using pointing gestures to actively direct the attention of the robot's caregiver, resulting in learning times that are two orders of magnitude less than comparable published models and eliminating the need for hand-labeling training data. A dedicated system to evaluate this idea is presented and its performance is demonstrated.

From Motor Learning to Social Learning: A Study of Development on a Humanoid Robot

A Dissertation

Presented to the Faculty of the Graduate School

of

Yale University

in Candidacy for the Degree of

Doctor of Philosophy

by

Ganghua Sun

Thesis Advisor: Professor Brian Scassellati

July, 2006

Contents

Chapter 1

Introduction

1.1 Motivation

As a scientific and engineering discipline, humanoid robotics has advanced significantly in the past few decades. Today's state-of-art humanoid robots such as ASIMO and QRIO are far cries from the heavy and clumsy first-generation robots.

There are two driving forces behind the growth of humanoid robotics. The first one is the hope that one day humanoid robots will be able to perform useful tasks for society, e.g. assisting the elderly or working in hazardous environments. The second force is the use of humanoid robots as experimental platforms to test ideas about artificial intelligence. No matter what the ultimate goal is, it has been recognized that engineering alone does not scale up to the challenges ahead, which, according to Brooks et al.[23] and Asada et al.[5], can only be met by allowing humanoid robots to develop and learn just as infants do. In recent years, developmental learning as a methodology has been quickly adopted by researchers not only in humanoid robotics but also in other fields of robotics. (See [86] for a summary of the applications of this field.)

Despite some success, the full potential of developmental learning is yet to be realized. In particular, one important feature of development — skill progression —

has been largely ignored in humanoid robotics. Skill progression refers to the fact that although there is a great deal of variability among individual infants, typical skills in different domains are acquired in an orderly fashion. Infants usually (but not always) learn to sit and crawl before they start to walk. Single words are uttered before syntax and grammar are mastered. Such a progressive development of skills allows for a structured decomposition of a complex system. Constraints based on limited perceptual or cognitive capabilities of an infant aid in the acquisition of complex skills by allowing learning to occur first in lower-dimensional spaces. As infants master basic skills, the already acquired skills become useful tools to reduce the complexity of learning more complex skills.

Humanoid robotics started as an engineering discipline and many researchers in the field still tend to rely on "engineering thinking" to solve problems. It is a common practice in engineering to break a complex problem into smaller pieces and then focus on the smaller pieces one at a time. When this divide-and-conquer approach is directly applied to skill learning on a humanoid robot, it often causes one to focus too much on individual skills and to overlook the connections between them. When the learning of an advanced skill is taken out of the development context, it often appears to be a much harder problem (e.g., learning in a very high dimensional space) that warrants a large number of training samples or the application of exotic techniques. The complexity of the problem could, however, be substantially reduced if the robot were able to take advantage of the benefits of skill progression and learn the underlying and more basic skills first.

1.2 Overview

The primary contribution of this thesis is that it provides the first demonstration of the efficiency of skill progression in the field of humanoid robotics. The demonstration uses Nico, an upper-torso humanoid robot designed and constructed by the author over the past few years, as the experiment platform. Nico's body dimensions and kinematic structure closely match those of an average one-year-old such that Nico faces similar physical constraints to one-year-old infants. The small body size of Nico also tends to make experiment subjects treat and respond to Nico as if it were an infant. This thesis describes how Nico acquires reaching, pointing and joint attention — three important skills that span the domains of the motor learning and social learning. Although the close connections among these skills have long been recognized in developmental psychology, this thesis represents the first attempt to exploit these connections for building computational models of these skills in an incremental fashion. To improve the performance of the models, various aspects of developmental learning have been extensively examined and explored.

The development of reaching is based on motor babbling, which is driven by the self-exploration instinct of the robot. The production of pointing gestures is developed out of the reaching skill and an interplay between the body structure of the robot and the cooperative nature of its caregiver. After the development of pointing, the robot then actively uses pointing gestures to manipulate its caregiver's gaze direction for the learning of joint attention. This progression of skill development allows the robot to acquire the joint attention skill with a training set that is a few magnitudes smaller than that required by previous models. Throughout the development process, only a common type of neural network — Radial Basis Function Networks (RBFN) — has been used. This on the one hand demonstrates the power of skill progression that eliminates the need for highly customized learning algorithms. On the other hand, it seems to provide support to the conjecture raised by many developmental psychologists that only a few general learning mechanisms are required for the acquisition of a large variety of functions/skills.

The thesis is organized as follows:

- *Chapter 2*: Background

  In the first part of this chapter various facets of development psychology that are relevant to the main body of the thesis are described. In particular, findings regarding the development of reaching, pointing and joint attention are presented. In the second part of this chapter, previous work in humanoid robotics on skill learning is introduced.

- *Chapter 3*: Nico: A New Humanoid Robot

  The main features of both the hardware and the software system of Nico, a new humanoid robot that served as the experiment platform for this thesis, are presented.

- *Chapter 4*: Learning to Reach

  After an examination of previous models, a new model for generating reaching movements toward spatial targets is proposed. It is based on a learned model of the forward kinematics of the arm and a directional mapping that is extracted from the forward model. These two components enable the robot to reach targets with or without visual feedback.

- *Chapter 5*: Extensions of the Reaching Model

  Two extensions of the reaching model are presented. The first extension incorporates the neck joints of the robot into the forward model without increasing the dimensions of the learning space. This results in gently curved reaching trajectories similar to those produced by humans. The second extension makes a minor modification of the reaching model to handle the situation of kinematic singularity. This naturally produces a pointing gesture when the target it tries to reach is too far away.

- *Chapter 6*: Active Learning of Joint Attention

  The concept and the details of the experiments for joint attention learning are described. In contrast to previous approaches, the experiments exploit the active nature of infants and the social dynamics between infants and their caregivers. It is demonstrated that by actively pointing and watching for contingent responses, Nico is able to learn the joint attention skill online through its interactions with multiple experiment subjects.

- *Chapter 7*: Conclusion

  A summary of the contributions of the thesis is given. The results from the previous chapters are re-evaluated in terms of the performance of both the individual pieces and the overall system.

Chapter 2

Background

This chapter provides background on the two fields that are most relevant to this thesis: developmental psychology and humanoid robotics. First, an introduction of the major theories of developmental psychology is provided. The current status on the nature versus nurture debate is summarized and philosophy of the constructivist view of development described there serves as an important guide for this thesis. Next, two concrete examples drawn from developmental psychology are used to illustrate that in contrast to divide-and-conquer approach commonly used in engineering, development is a dynamic process where the existence of one skill can greatly benefit the acquisition of another skill. Following these examples, an important learning mechanism that forms the basis of a later chapter in this thesis is described and details regarding the development of reaching, pointing and joint attention in infants are presented.

The second section of this chapter provides a brief history of humanoid robots. Particular emphasis is placed on the learning of perceptual and motor skills and on various aspects of social skills. The relationship of other humanoid robot projects to this thesis is discussed at the end of the chapter.

2.1 Developmental Psychology

2.1.1 Influential Theories of Development

According to [102], developmental psychology can be defined as "the discipline that attempts to describe and explain the changes that occur over time in the thought, behavior, reasoning, and functioning of a person due to biological, individual, and environmental influences". Its origin can be traced back to Rene Descartes and John Locke. Among the most influential modern day theories are Piaget's cognitive-stage theory, Gibson's perceptual-development theory, social learning theory and connectionist theory[96].

- Piaget's theory posits that children go through four distinct stages of development as they grow and that they continually build new mental representations or *schemas* of the world. Learning takes place through *assimilation* (fitting the reality into the current schemas) and *accomodation* (modifying schemas to adapt to new experiences).

- In contrast to Piaget's view that children construct over time more elaborate representations of the world, Gibson maintains that development is a process during which children become more efficient at extracting information from perception [56]. The concept of *affordance* occupies a special position in Gibson's theory. An object's *affordances* are the actions offered by the environment and the object and is believed to be directly perceived by humans [57]. It intimately links perception and action together.

- According to Albert Bandura, the pioneer of modern social learning theory, the confluence of past experiences and internal preferences makes each child behave in a unique way [7]. The environment responds to different behaviors in different ways and this further leads to divergent experiences. In this sense, children *create* their environment.

- Over the past few decades, psychologists have increasingly used the metaphor of the human mind as an information processing machine. The advent of artificial neural networks (ANNs) provides researchers with new tools to represent mental functions of humans with quantitative models. Models based on ANNs are often called *connectionist* models. Such models often take perceptual information as their input. Their output can either lead to observable behaviors or be routed to higher-level models.

Each of the four theories very briefly described above offers an overarching framework for developmental psychologists. Although contemporary developmental psychologists often study development from their own unique perspectives, these theories remain influential. A common theme of these theories is the recognition that children play an active role during their development. According to Piaget, children actively reconcile their mental representations of the world with the reality through assimilation and accommodation. Like Piaget, Gibson views children as inherently motivated creatures that try to understand the world with active exploration. Bandura's social learning theory gives the activeness of children a prominent place as they create their environment. Though connectionist theory is more focused on replicating mental functions with connectionist models, the actions of children heavily influence the input to these models and consequently the update of their internal parameters.

From birth to the age of around 24 months, infants change from helpless creatures that are unable to hold up their head to active explorers that are perpetually on the move, speak a few hundred words and interact socially with their parents and peers. Studies of infant development focus on the changes taking placing during this

Domain	Research topics and representative works
Perceptual development	Face perception [61], depth perception [58], perception of biological motions [52] and auditory tuning to native language [145].
Motor development	Reaching [141], grasping [110], pointing [135], crawling [1], walking [131] and posture control [12].
Cognitive development	A-not-B effect [133], numerical concepts [151] and object categorization [118].
Social development	Imitation [90], joint attention [121], theory of mind [8] and language acquisition [17].

Table 2.1: Major domains of infant development studies

important period of time. Researchers believe that by studying the development of infants, a deeper understanding of the behaviors displayed by older children and adults can be achieved. They also hope that with a detailed characterization of normal development available, abnormal development can be detected at the earliest sign so that clinical intervention can be introduced as soon as possible.

The body of infant development research consists of the following four major domains: perceptual development, motor development, cognitive development and social development. Several important research topics along with the most representative works for each domain are listed in Table 2.1. It should be noted that some of the topics in the table can be categorized into multiple domains. For example, pointing is often studied within the context of social development because of its referential purpose. The development of reaching, pointing and joint attention in infants will be described in detail in Section 2.1.5.

2.1.2 Nature versus Nurture

For centuries, philosophers and scientists have been debating what makes us become who we are. Is it nature or is it nurture? At the one end of the continuum is Rene Descartes's belief that human behaviors are fully determined by the innate resources of the mind. At the other end is John Locke's *tabula rasa* view that humans come to the world with a mind like a blank slate that awaits to be written on by experience.

Today the *tabula rasa* view in its pure form is largely abandoned since the effect of genes on a living organism is simply undeniable. On the other hand, the landscape of development psychology is stilled dotted with strong nativist rhetoric. In his book *Modularity of Mind* published in 1983, philosopher Jerry A. Fodor presents a view that the mind is composed of domain specific modules. According to this view, development is a maturation process of these modules rather than a learning process [51].

Although Fodor's radical nativism has been challenged by philosophers of diverse orientations, its influence can be still felt. For example, some researchers have suggested that face perception is innate. This hypothesis is supported by experiments showing that newborn infants exhibit a preference to face-like stimuli and are able to imitate the expressions on the face they see [61][90]. Additional support is provided by clinical studies of prosopagnosia (face blindness), which show that face perception in adults are localized in a specific brain region [44][9]. Another example of nativism is the "theory of mind" proposed by Baron-Cohen who suggests that a innate system consisting of four specialized modules provides the neural basis for humans' social behaviors [8]. Among the four modules is an Eye-Direction Detector (EDD) that gives people the ability to determine where their communication partners are looking at.

In contrast to the nativist/maturation views on development, another line of thinking, often labeled as *constructivism*, suggests that for many of the brain functions, genes have only specified some general neural structures and development mechanisms rather than containing detailed blueprints [76][72]. Recent experiments show that many seemingly sophisticated functions indeed do not have to be directly implemented by genetic programs. They can either emerge as the aggregates of several interconnected general structures or be based on general structures that are subsequently molded by experiences. The constructivist explanations on the origins of face perception and eye direction detection are described below.

Several cleverly designed experiments carried out by Francesca Simion et al. show that newborn infants show a preference to top-heavy geometric arrangements of square elements and this preference can be explained by the sensitivity difference between the upper and the lower visual field [128]. Since the elements of humans faces are also arranged in a top-heavy way, Simion et al. argue that humans' expertise in face processing may be the result of the interaction between a bias in the visual system and postnatal experiences. In [55], Gauthier et al. point out that the fusiform gyrus, which is often seen as the innate cortical module for face perception, also shows increased activity in individuals who have recently acquired expertise in recognizing artificial objects called greebles. This discovery provides support for the view that face perception is based on a more general neural structure for object perception.

The infant experiments carried out by Mark Johnson et al. cast doubt on the existence of an innate brain module for eye direction detection [45]. They demonstrate that infants' perception of eye direction is heavily influenced by motion cues and the presence of a face. They propose that the function of eye direction detection can be imbedded in the general structures for motion detection and face perception. Face perception, in turn, can be developed out of a bias in the visual system and postnatal experiences as has just been argued in the last paragraph. In [37], Corkum and Moore show that infants younger than 15 months mostly rely on head orientation instead of eye orientation to assess the attention focus of their caregivers. This result implies that a dedicated eye direction detector does not have to exist at birth; it can be developed from postnatal experiences.

2.1.3 The Dynamics of Development

Divide-and-conquer is one of the most important paradigms in engineering. It breaks a complex problem into a set of simpler sub-problems. The sub-problems are attacked individually and their solutions are later put together (typically in a sequential chain) to produce a solution to the larger problem. In contrast to divide-and-conquer, which is more or less static, development is a dynamic process. During this process multiple skills are learned in parallel and they exert influences over each other. For a better appreciation of the dynamic aspect of development, two concrete examples are presented below.

Co-development of locomotion and perception of optical flow [4]

Visual perception plays an important role in postural control. When the body travels through space, the visual system determines the positional and orientational change of the body with respect to the environment using optical flow information. It has been discovered that young infants use both central foveal vision and peripheral vision for postural control while adults rely primarily on the peripheral vision for this purpose. This phenomenon gives rise to the question of when infants start to switch to a more economical way of using only peripheral vision for postural control. Experiments show that the transition occurs during the time infants start to learn locomotion skills such as crawling. But why should the development of these two skills coincide? The answer lies in the different patterns of optical flow the vision system encounters under different conditions. When infants are passively transported, the angle between their line of sight and their moving direction is arbitrary. So the patterns of optical flow falling on the central visual field and the peripheral visual field are unpredictable, which makes them more difficult to decode. When infants start to crawl, their line of sight and their movement direction usually overlap. This makes radial optical flow and lamellar optical flow fall on the central visual field and the peripheral visual field respectively. Over time, the peripheral visual field becomes specialized to lamellar optical flow. The central visual field gradually focuses more on other tasks such as steering around obstacles and detecting surface conditions. This in turn enhances infants' locomotion skills.

Joint attention facilitates word learning [6]

Word learning is a difficult task. A simplistic model for word learning would suggest that infants could learn the meaning of new words simply by detecting the co-occurrences between the words their parents say and the objects in their environment. But a careful examination of this model reveals that such a covariance detection process would result in many mapping errors. For example, an infant can be playing with a toy when his/her mother is making a phone call. In this scenario, the words the infant hears bear no relationship to the toy the infant is focussing on. However, studies on infant word learning show that infants scarcely acquire false mappings for concrete nouns. How do infants manage to only learn correct mappings most of the time? Experiments conducted by Baldwin show that 18-month-old infants rely on joint attention skill to learn new words. When infants at this age hear a new word, they actively look back at the speaker and only associate the new word to the object at the speaker's attention focus. This mechanism leads to a correct association of the new word and the object it refers to. Baldwin conjectures that the extraordinary pace at which infants at 18-20 months acquire new words can be attributed to a well-developed joint attention skill.

The first example demonstrates that locomotion aids the visual system in processing optical flow information more efficiently such that posture control can be achieved using only peripheral vision. On the other hand, better posture control and the released computational resources for central vision facilitate the learning of locomotion. The second example shows that a well-developed joint attention skill helps disambiguate referents in the environment to make word learning possible. Locomotion skills and the processing of optical flow could probably be engineered separately. However, to design an artificial system that can learn new words in a natural environment without engaging joint attention skill would be extremely difficult, if possible at all.

2.1.4 Contingency Learning

Contingency is the relationship between behavior and its consequences. In a landmark paper, Murray and Trevarthen [104] used a closed-circuit video system to study infants' sensitivity to social contingency. In the initial experiment, mothers interact contingently with their infants through the video system. The infants respond positively to the contingent interactions by gazing more at the video and smiling. In the second experiment, the live interactions recorded during the first experiment are replayed to the infants. The infants become visibly upset in the absence of contingency between infants' own actions and mothers' actions in the video. This experiment paradigm has been rapidly adopted by other researchers whose studies confirm infants' ability to detect contingency [63][106].

Other studies have used objects such as a furry animal [73] or a simple robot [101] to investigate whether novel objects can invoke similar responses of infants by displaying contingent behaviors. The outcome of these studies demonstrates that contingency is a powerful mechanism that affects infants' attitude toward both people and objects. Infants even exhibit gaze following behavior during their interactions with a box-like robot (Baby-9) that does not bear any resemblance to a human [101]. This observation is remarkable because some researchers believe that gaze following is based on an innate eye-direction detection module (see Section 2.1.2). It suggests that contingency may play a critical role in the development of other functions. For example, it has been postulated that contingency detection can be used by infants to distinguish self from others because of the difference between self contingency and social contingency [143][144]. An example of self contingency is seeing one's own body part move after sending a motor command while an example of social contingency is seeing other people react to one's own action.

After publishing the experiments on infants' response to Baby-9, Movellan designed Infomax, a sophisticated model based on information theory for the detection of social contingency [100]. Infomax was recently used on a robot that can spontaneously make noises through a mounted microphone [27]. After it makes a noise, the robot captures an image from its camera and labels the image as positive when Infomax detects a contingent response. These data are then used to train an object classifier. After the training, the classifier shows a remarkable preference for face-like stimuli because the images labeled as positive usually contain a human face. This result again shows that contingency can be used as a powerful mechanism for an infant to learn from its environment.

Movellan's experiments show that his contingency detector Infomax has many potential uses. However, Infomax is a complex model. Evidences from infant studies suggest that a simpler contigency detector based on time delay may benefit learning as well. Early studies show that infants learn the association between their action and the outcome only when the the outcome occurs within a 3 second time window after infants' action [95]. More recent results by Striano, Henning and Stahl (unpublished) show that infants are even sensitive to when the outcome occurs within this already very narrow time window. Michel, Gold and Scassellati recently report that by relying on a simple model of the time delay between the output of motor commands and the perception of the motion of its own end effector, a humanoid robot can pass the mirror test of self-recognition [111][54].

2.1.5 Development of Reaching, Pointing and Joint Attention

Reaching

According to some experiments, newborn infants can already extend their arms in the general direction of a target when they are properly supported [140][3]. Longitudinal studies by von Hofsten, Thelen and Konczak show that the first time an infant succeeds in contacting a stationary target after an attempted reach occurs at an age of three to five months [141][132][77]. Three months after the onset of successful reaches, infants can already reliably reach toward spatial targets although adult-like reaching movements are only seen after the age of two years.

Von Hofsten reported that young infants' reaching movements consist of multiple segments [141]. Each of the segments has an acceleration phase and a deceleration phase. While von Hofsten found that the number of the segments in each reaching movement decreases dramatically when infants grow older and become more experienced with reaching, other researchers found that the number of segments decreases only moderately or stays stable [46][132][15]. This difference in opinion may the result of the different ways movement data are filtered. This controversy aside, it is generally agreed upon that infants' reaching trajectories are significantly curved and become straighter over time.

Initially, it was believed that infants continually monitor the position of their hand visually and correct its moving direction until it reaches the target. The multiple segments in their reaching trajectories would then reflect such online corrections. However, Clifton and her colleagues later discovered that infants can reach toward glowing objects in the dark [35]. This discovery means that infants do not necessarily need to visually monitor their hand while they reach. It suggests that infants can also use proprioception to guide their hand toward a target when visual feedback is not available.

Although adults can apparently reach toward a target without much effort, the control behind reaching movements are by no means simple. At the kinematic level, there are 7 degrees of freedom in the human arm that can be simultaneously controlled. If this does not seem to be complicated enough, humans can only exert direct control over the muscles that generate the forces to make the arm move. The number of degrees of freedom in the arm at the force control level is much larger than that at the kinematics level. Bernstein was the first to point out this "degrees of freedom problem" [11].

Bernstein believed that well coordinated arm movements are unlikely to emerge without a sound learning strategy. He proposed a three-stage process of motor learning based on his observations. According to Bernstein, when a human learns a new motor skill, he/she first freezes his/her distal joints (joints at the periphery) to reduce the dimensionality of the control space. Later, the distal joints are gradually released and incorporated into the newly acquired skill. At the end, reactive phenomena such as gravity are exploited to refine the skill. Data of several studies on adults' learning of new motor skills such as dart throwing and skiing are supportive of Berstein's proposition [89][139]. Other researchers suggest that during the time infants learn the reaching skill, they may just be forced by nature to follow the three-stage process proposed by Bernstein [13][14]; due to the proximodistal direction of maturation of the neural and muscular system supporting arm movements (joints close to the body develop before joints close to the hand), infants mostly use their shoulder joints to move the hand around at the time they start to reach toward spatial targets; the more distal joints in the elbow and the wrist are kept stiff. Only later are these joints used for reaching when their control system has matured.

Another proposed mechanism to harness the dimensionality of the control space during motor learning is based on the theory of muscle synergy. This theory posits that the base control unit of the central nervous system (CNS) is muscle synergy, sometimes also called motor primitives, instead of individual muscles. A muscle synergy is a simultaneous activation of multiple muscles that leads to a distinct limb posture. Muscle synergy was first discovered in frogs [60]; and it seems that humans use muscle synergies for the control of arm movements as well[124]. These observations have led to the hypothesis that each arm movement corresponds to a trajectory in a virtual space where each point is a linear combination of muscle synergies [105]. The task of CNS is to activate the right synergies at the right time. Muscle synergies shield the CNS from the nasty nonlinearity of direct muscle control. Because the number of muscle synergies is believed to be small in comparison with the number of muscles, the dimensionality of the control space is reduced, thus making motor learning easier.

Pointing

Pointing is often seen as a communicative gesture. A canonical point is defined by Butterworth as a posture with both the index finger and the arm extended in the direction of an object [29]. Vygotsky suggests that pointing is simply an adapted form of reaching [142]. According to Vygotsky, during the practice of reaching toward attractive objects in the environment, an infant occasionally fails to reach an object because it is too far away. Such a failed reach has the characteristics of a fully extended arm toward the desired object. The caregiver sees this situation and retrieves the object for the infant. Through repetition of this scenario, the infant learns that showing the posture of an extended arm toward an object can serve as a means to obtain it. This way, a failed reach is gradually transformed into a prototype of pointing that the infant uses as an instrument to obtain distal objects. Over time, the infant observes the canonical points displayed by the mother and through imitation, increasingly points in the canonical way.

An intriguing study by Povinelli and Davis shows that in anesthetized humans, the index finger of an arm at rest is always sticking out with respect to other fingers due to the anatomy of the human arm [116]. So even before a young infant adopts the canonical pointing gesture, the index finger may be inadvertently extended for anatomical reasons. Vygotsky's account on the origin of pointing implicates that infants do not initially understand the functional significance of pointing. This is supported by experiments showing that early pointing also occurs without the presence of adults and that young infants show both reaching and pointing behaviors toward distal objects [10][103].

According to Butterworth, the average age of the onset of canonical pointing is

11 months [29]. If Vygotsky's theory on the origin of pointing is true, the onset of pointing can be dated back several months earlier when infants can already reliably reach close objects. Butterworth, however, does not agree with Vygotsky's theory. He believes that canonical pointing is a human-specific adaption. Butterworth and his colleagues have presented evidence suggesting that canonical pointing does not develop out of reaching, but serves the purpose of establishing joint attention first its first onset [53]. In particular, they argue that canonical/index-finger pointing may have evolved from pincer grip, which is a human-specific way of grasping by closing the thumb and the index finger onto an object [31].

Although Butterworth's view on the origin of canonical pointing has several supporters [85][33], other researchers point out that canonical pointing is not a unique gesture for referential purpose in some non-western cultures; in those cultures, pointing with the whole hand is commonly used as well [67][147]. Most interestingly, David Wilkins mentions that Micheal Olson, an expert on the Barai culture of Papua New Guinea, found that Barai people do not understand the referential meaning of canonical/index-finger pointing [147]. These observations suggest that relying on canonical pointing to direct other people's attention may be a cultural phenomenon rather than a universal phenomenon.

Joint Attention

Joint attention, sometimes called joint visual attention, can be defined operationally as "looking where someone else is looking". It also can be defined more subtly as following the direction of attention of another person to the object of their attention [43]. Joint attention is not merely a coincidence of two lines of gaze; it serves a number of very important functions in the development of an infant. For example, by assessing the mother's attention focus and her facial expression, infants can determine whether a novel object is dangerous [129]. Joint attention can also help infants learn the correct mapping from a word spoken by the caregiver to an associated object [6][17]. It has been suggested that a deficit of joint attention can lead to serious disorders such as autism [82][8].

A paper published by Scaife and Bruner in 1975 is often cited as one of the most influential studies on joint attention [121]. In that paper, Scaife and Bruner described an experiment on thirty-four infants ranging in age from 2 to 14 months. During an experimental trial, an infants sat facing an experimenter. After some initial interactions, the experimenter made two head turns, one 90° to the right and one 90° to the left, with a interval of 20 to 50 seconds between the two turns. When the experimenter made a head turn 30% of 2 month old infants looked in the same direction. Infants' responses during the experiment were recorded. Scaife and Bruner found that 30% of 2 month old infants looked to the same direction when the experimenter made a head turn. The percentage grew quickly with increasing age. Scaife and Bruner concluded from this experiment that infants had from a very early age a rudimentary ability to follow other people's attention.

Later, Butterworth and Jarrett detailed a three stage model for the development of joint attention in infants [28]. The first stage corresponds to an age of around 6 months. At this stage, infants are able to turn their head in the same direction as their mother, but will stop at the first object they encounter. The second stage corresponds to the age of 12 months. At this stage, infants are not only able to determine the direction their mother is looking, but also find the right object if it is initially within their field of view. At the third stage, or at the age of 18 months, infants are eventually able to find the object their mother is looking at even when it is initially out of view.

Although both the work of Scaife and Brunner and the work of Butterworth and Jarrett report that a significant proportion of infants at the age of 6 months can engage in joint attention [121][28], other researchers believe that the emergence of joint attention in infants does not occur until months later. The lack of agreement on when joint attention begins is partly caused by variations in experiment settings and evaluation criteria. For example, infants may spontaneously change their gaze direction [149]. When Corkum and Moore applied a more stringent criterion to study the onset of joint attention, they found that infants at the age of 12 months exhibit only very rudimentary joint attention skill and that only after the age of 15 months do infants show reliable joint attention responses characterized by significantly more matches of their gaze direction and the experimenter's gaze direction than mismatches [37].

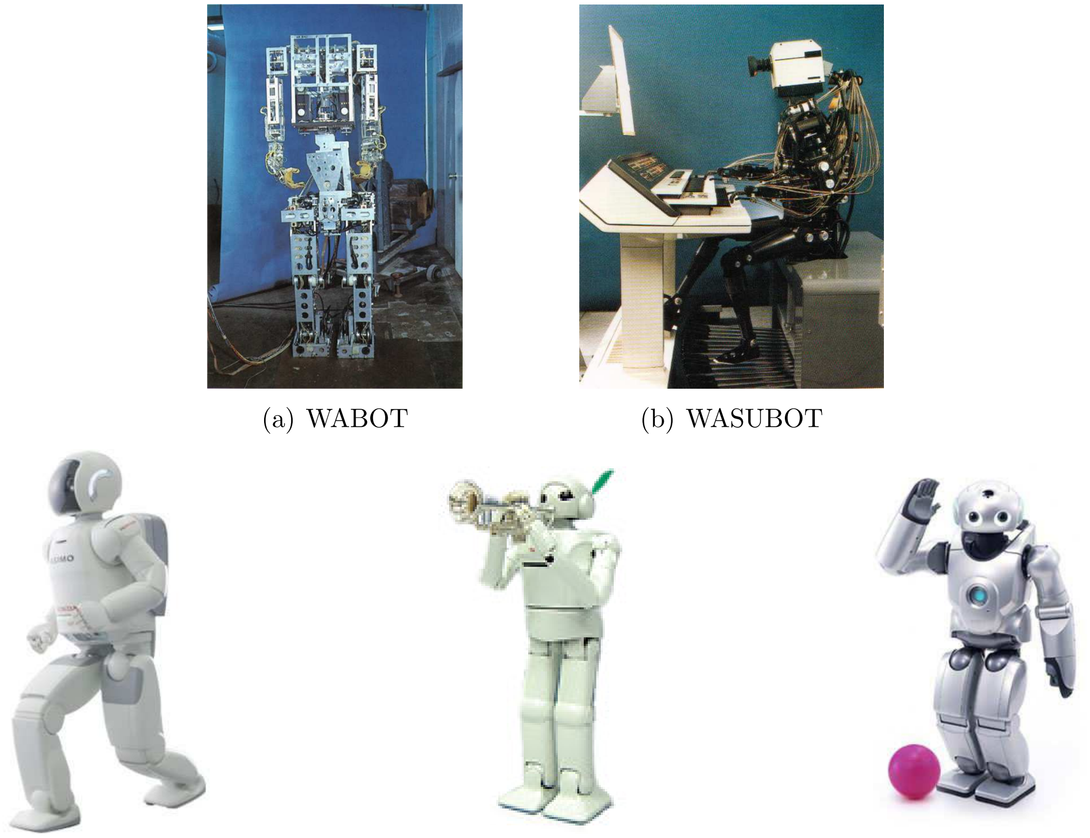
Besides the studies to determine when joint attention starts, researchers have also tried to determine how the performance of joint attention is influenced by the cues exhibited by the caregiver. As mentioned earlier, Corkum and Moore found that before the age of 15 months, infants mostly rely on the caregiver's head orientation to assess his/her attention focus [37]. Interestingly, Butterworth and Itakura found that adults can locate a target more accurately when the experimenter wears sunglasses than when the experimenter's eyes are visible [30]. In parallel to the studies on infants' response to the change of caregiver's head or eye orientation, experiments have been conducted to examine infants' comprehension of pointing gestures. Murphy and Messer have observed that it often takes significant effort for an adult to make young infants follow a pointing hand to the object [103]. Morissette and his colleagues also found that infants younger than 12 months mostly stare at the pointing hand rather than follow its direction [99]. Messer believes that infants' comprehension of pointing gestures before the age of 12 months is opportunistic at best [91]. This evidence suggests that pointing comprehension emerges later than pointing production.

2.2 Humanoid Robotics

2.2.1 A Brief History of Humanoid Robots

The term "humanoid robot" loosely refers to robots with a morphology similar to that of humans. These robots usually possess a subset of the sensory, motor and cognitive functions of humans.

Japan has a longer history of building humanoid robots than any other country in the world. The long-term goal of many humanoid projects in Japan is to build robots that can be integrated into society and serve useful roles such as caring for the elderly. Although the humanoids built in Japan are usually equipped with both a vision system and a communication system, their sophisticated motor skills are what truly distinguish them. Often considered as the first full-scale humanoid robot in history, WABOT (Fig. 2.1(a)) was constructed under the direction of Ichiro Kato at Waseda University in 1973. It was able to walk, although in an awkward fashion, and transport objects with its gripper. These simple acts were no small achievement at that time. However, it was WASUBOT (Fig. 2.1(b)), another creation of Ichiro Kato constructed more than a decade later, that truly stunned the world. At its debut at Japan's Expo '85, WASUBOT played J.S. Bach's *Air on the G String* on a keyboard, accompanied by the NHK Symphony Orchestra. Today, more than three decades after the birth of WABOT, Japan is home to the largest proportion of the world's humanoid robots. Among these are Honda's ASIMO (Fig. 2.1(c)), Toyota's Partner Robots (Fig. 2.1(d)) and Sony's QRIO (Fig. 2.1(e)). All of these robots possess a large number of degrees of freedom and can execute complex motor skills such as



(a) WABOT

(b) WASUBOT

(c) ASIMO

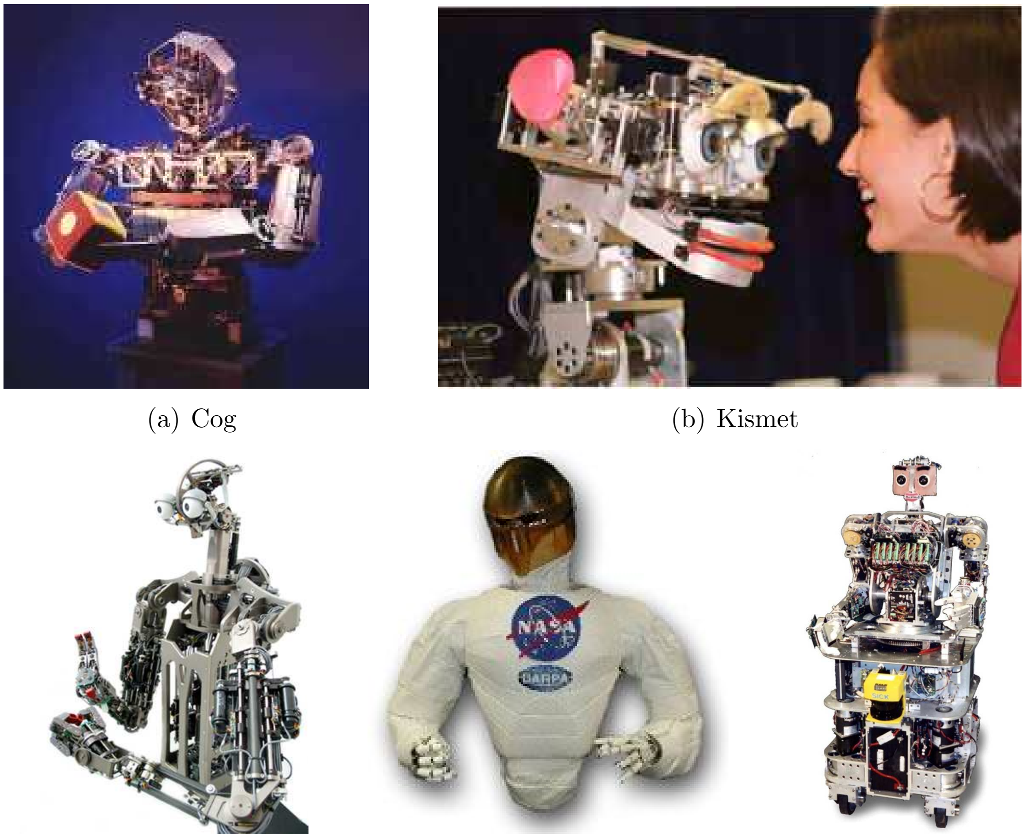
(d) Partner Robot

(e) QRIO

Figure 2.1: Humanoid robots designed in Japan

climbing stairs, playing the trumpet and dancing.

The history of humanoid robots in the United States started out with a few scattered projects such as the Green Man and the Robotic Mannequin that were designed specifically for military studies. In 1993, the MIT AI lab started the construction of a humanoid robot called Cog (Fig. 2.2(a)) under the direction of Rodney Brooks. The goal of the Cog project was to use a humanoid robot as a platform to study human intelligence. Achieving human level intelligence has always been the dream of researchers in the field of Artificial Intelligence (AI). In the three decades following the establishment of this field in the 50s, most work in AI was carried out based on the assumptions that perfect representations of the world are available and that



(a) Cog

(b) Kismet

(c) Domo

(d) Robonaut

(e) Dav

Figure 2.2: Humanoid robots designed in the United States

intelligence manifests itself mainly in the decision making or problem solving processes based on these representations. Brooks suggested that true intelligence could only emerge from a physical body situated in the real world [22]. Building on his early successes with insect-like robots, Brooks described a grand plan for Cog [24]. Although some of the most ambitious parts of that plan were eventually eventually not realized, the Cog project attracted the interest of many researchers in humanoid robots and spawned several other humanoid projects such as Kismet (Fig. 2.2(b)), and most recently, Domo (Fig. 2.2(c)) [19][42]. Other humanoid robots built in US include Robonaut (Fig. 2.2(d)) by NASA and Dav (Fig. 2.2(e)) by Michigan State University [2][64].

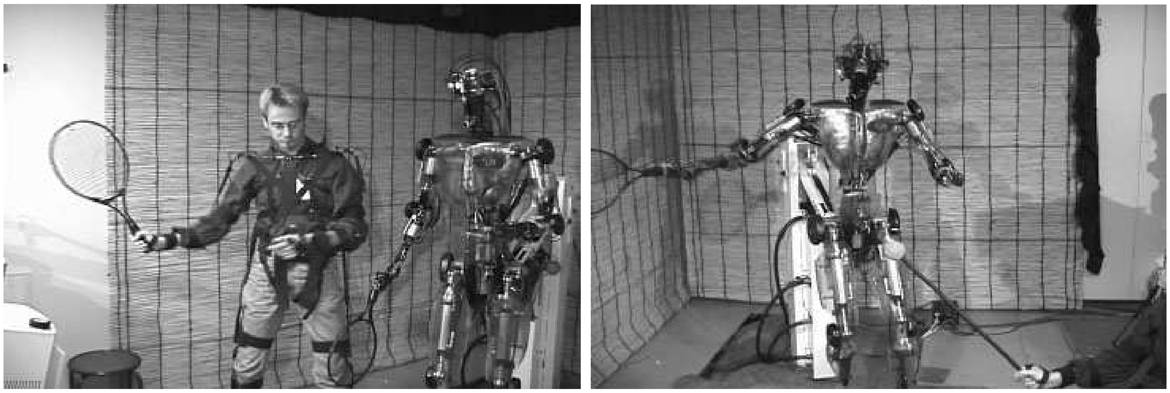
2.2.2 Sensory and Motor Learning

Imitation Learning

Imitation is the ability to recognize and reproduce others' actions. While the controversy of whether infants can imitate other people immediately after they are born is not yet settled [90], there is no doubt that imitation learning is essential to both children and adults for the acquisition of a variety of skills. Ethological studies have revealed that imitation learning exist in other species as well. Neuroscience studies show that a neural circuit in the brain of monkey becomes active both when the monkey sees another monkey or a human manipulate an object and when it performs the same manipulation [120]. Similar neural circuits have also been found in humans [38]. These circuits are often called the "mirror neuron" systems and are believed to serve a critical role in imitation learning.

Most research on imitation learning in humanoid robots focuses on learning of skillful movements. Ideally, the information regarding these movements is delivered by a vision system. However, current computer vision systems are still unable to reliably parse human movements. Typically, when a person demonstrates a movement to a humanoid robot, several motion sensors are attached to his/her body to gather data regarding this movement. Sometimes, the demonstrator also puts a colored marker on his hand so that the position of the hand can be continuously monitored by a dedicated vision system. Such experimental settings produce a large amount of raw data from even a single demonstration. Thus, it is important to find an efficient representation of these data.

Inspired by the discovery that humans use muscle synergies/motor primitives to simplify the control of their limbs, Auke Ijspeer and his colleagues suggested that movements can be decomposed into a combination of primitives as well [71]. They

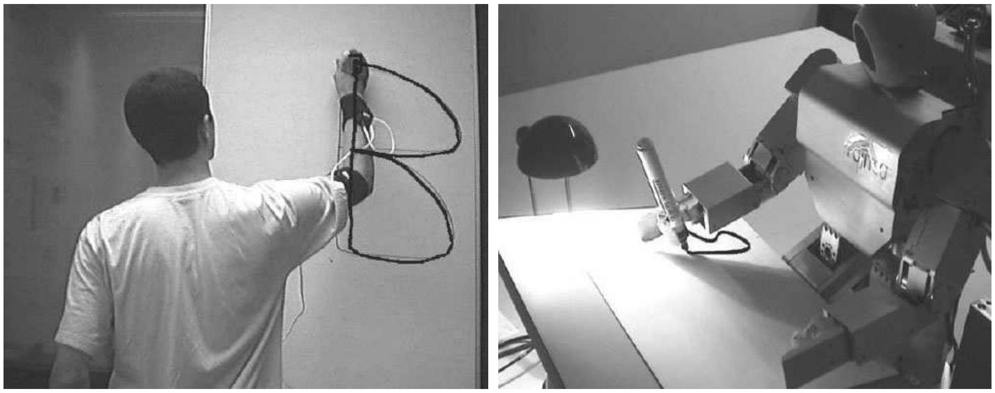


(a) (b)

Figure 2.3: Learning tennis swing by imitation [71].

showed that by using attractor dynamics or limit cycle dynamics (both of which can be described with autonomous differential equations) as primitives, complicated movements such as a tennis swing (see Fig. 2.3) can be reliably reproduced on humanoid robots. The usefulness of using differential equations as movement primitives has also been demonstrated by a study on learning biped locomotion by imitation [109]. The research group led by Maja Mataric demonstrated that instead of being designed by hand, movement primitives can be automatically derived from human movement data as well [50].

Calinon and Billard's work on gesture learning is another study on imitation learning [32]. They use both motion sensors and a vision system to gather data from the demonstration of several gestures. The resulting data consist of a sequence of joint angle vector and hand position vector. These data are first preprocessed by PCA to produce a more compact representation. Then they are fed into a learning algorithm to construct a Hidden Markov Model (HMM) for each of the gestures. Recognition of a new gesture can be achieved by estimating the likelihood that the newly observed gesture is generated by one of the learned HMMs. By recovering the joint vector sequence from a learned HMM and interpolating with splines, the



(a) (b)

Figure 2.4: Learning gesture by imitation [32].

corresponding gesture can be reproduced on a humanoid robot (see Fig. 2.4).

Progressive Learning

In his Ph.D. thesis, Giorgio Metta gives a description of a simulated development of motor skills on Babybot (see Fig. 2.5), a humanoid robot developed by the Lira-Lab at the University of Genoa [92]. The stages of development for Babybot are organized in a way similar to how infants develop motor skills.

The first step of Babybot's development is to learn the closed-loop gain for target foveation, which relies on a feedback controller to incrementally move a target closer to the center of the visual field. Since such a feedback control scheme may take multiple steps to achieve foveation, a saccade map is learned as the next step. The saccade map converts the retinal error (the distance between the projection of a target and the center of visual field) to eye motor commands for adjusting gaze direction. During the third stage of the development, another map is learned for the generation of neck movement to orient the head towards a target. The eye control and the neck control operate independently in Babybot. These two redundant controls are mediated by a controller similar to the human Vestibular Ocular Reflex (VOR). This controller is

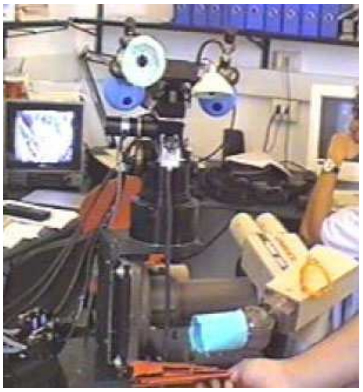


Figure 2.5: Babybot - the humanoid robot used by Metta for his experiments on motor skill development [92].

tuned in parallel to the learning of the closed loop gain, the saccade map and the neck map. The learning of all these control modules takes place through self-exploration and enables Babybot to efficiently adjust its gaze direction toward spatial targets.

Once the ability to control gaze direction is obtained, Babybot also develops the skill of reaching. The reaching skill is based on the motor primitive hypothesis proposed by Mussa-Ivaldi and Bizzi [105]. It is realized by constructing a head-arm map that converts the angles of the neck and eye joints to the coefficients of four hand-designed motor primitives of the arm. When Babybot finds an interesting target, it first adjusts its gaze direction toward the target and then generates commands for the arm motors using the information provided by the head-arm map.

In contrast to Metta's work, which describes a development scheme that progresses from the learning of gaze direction control to the learning of reaching, Paul Fitzpatrick's work assumes that basic motor skills are already acquired and focuses on how these motor skills can benefit the development of perceptual skills [48]. He carries out his experiments on the humanoid Cog (see Fig. 2.2(a)).

Fitzpatrick demonstrates that by actively exploring with the arm, the task of segmenting objects lying on a table becomes much easier for Cog (see Fig. 2.6). Once

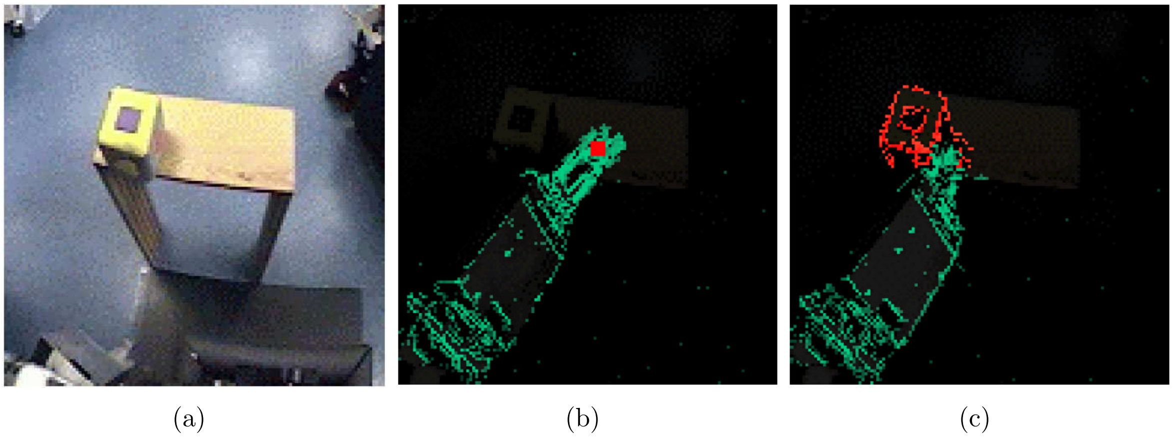


Figure 2.6: Discovering the boundary of an object by poking [48].

an object has been segmented, the orientation of its boundary can be determined. By decomposing the boundary into small patches and identifying the orientation of the edge within each patch, a lookup table can be constructed that maps small image patches to the orientation of the edges they contain. To test the performance of the lookup table constructed this way, Fitzpatrick applies it to both artificial and natural images and achieves remarkable results. He also shows that through manipulation, Cog can capture images of an object from different perspectives and build an appearance model of this object. Appearance models built in this way enable Cog to recognize an object.

2.2.3 Research on Social Skills for Humanoid Robots

Designing sociable robots

Designing sociable robots is a very challenging task. Many issues need to be resolved along the way. The two most important issues are (1) to make robots properly interpret the natural social cues exhibited by people and (2) to enable robots to produce these cues themselves. Additional issues include proper regulation of social interactions and real-time performance.

In [21], Breazeal describes in detail the design of Kismet, a sociable humanoid head known for its rich social behaviors (see Fig. 2.2(b)). Kismet is equipped with a variety of dedicated systems. Its facial expression system is powered by a large number of servo motors. There are nine basic facial postures such as fear, anger, surprise and disgust, which can be blended to create more nuanced expressions. Kismet's auditory system is designed to pick up the prosodic signals in human speech. These signals provide clues as to the emotional status of a person interacting with the robot. Kismet also has a vocalization system that can convey a repertoire of emotions by varying elements such as pitch, loudness and speech rate. Its vision system contains several low-level modules to process color, motion and etc. The outputs of these low-level modules are used by a higher-level module that decides what Kismet should pay attention to. A motivation system modifies Kismet's internal emotional status based on environmental stimuli. This internal emotional status in turn influences the way Kismet behaves toward the next stimulus. All these carefully engineered systems interact with each other and together ensure that Kismet can carry out natural social interactions with humans.

Several researchers believe that the social competency exhibited by humans can be to a great extent attributed to an innate brain module enabling a person to attribute desires, beliefs and goals to other people. Scassellati has designed a hybrid model for this hypothesized "mindreading" module [123]. Implemented piece by piece on the humanoid robot Cog (see Fig. 2.2(a)), his model combines the elements of the models proposed by Leslie and Baron-Cohen [83][8]. Scassellati shows that the high-level cognitive skills required by Leslie's and Baron-Cohen's theories can be realized with low-level perceptual abilities. The performance of his hybrid model is demonstrated on Cog in a basic social learning task.

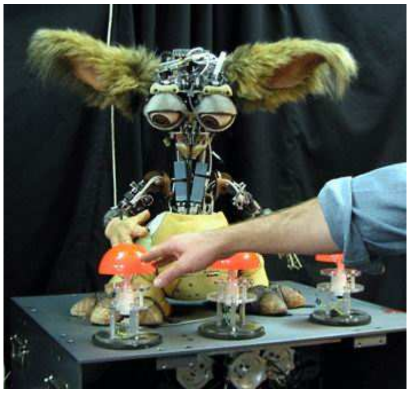
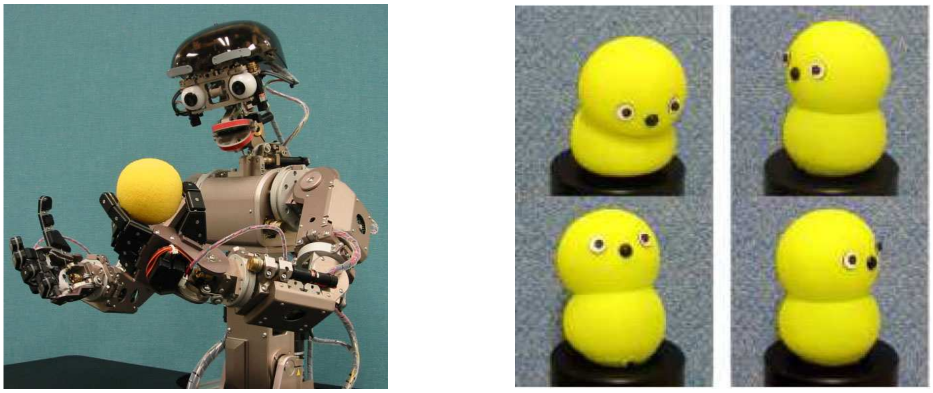


Figure 2.7: Leonardo learns to associate the rightmost button with a label by following the pointing gesture of the experimenter [20].

Leonardo is a new robot designed by Stan Winston Studio for the Robotic Life Group at the MIT Media Lab. Leonardo is not a humanoid robot in a strict sense; it has the appearance of a furry animal rather than a human. It is equipped with a sophisticated vision system and a speech recognition system. Its large number of degrees of freedom enable it to produce many social signals that are easy for humans to understand. The researchers in the Robotic Life Group use Leonardo as a platform to study human-robot collaboration. Extensive turn taking skills have been implemented on Leonardo such that the progress of the collaboration can be easily assessed [20]. For example, when an experimenter points to a novel object and assigns a label to it by speech, Leonardo will first follow the pointing gesture to the object and then look back to the person to acknowledge that an association between the object and the label has been established and further instructions can follow. In addition, Leonardo is programmed in such a way that it can build a hierarchical task representation through its interactions with a human.



(a) Infanoid

(b) Keepon

Figure 2.8: The two robots designed by Hideki Kozima to study the social development of children [79][81][80].

Using robots to study social development

Not only can sociable humanoid robots learn skills from humans in a natural setting, they can also serve as valuable tools to study human social behaviors. By tweaking the social modules inside these robots, different responses from humans may be observed. These observations can lead to a deeper understanding of the mechanisms that enable humans to behave socially. By carrying out longitudinal studies of how infants or children change their attitude toward a sociable robot put in their natural environment, a more detailed and accurate picture of when different social skills emerge and how they gradually improve can be depicted. More importantly, by investigating how children with developmental disorders differ in the way they interact with sociable robots, earlier diagnosis and intervention of these disorders may become possible.

Hideki Kozima and his team have carried out an experiment on how children of different ages interact with Infanoid, an upper-torso humanoid robot with basic social skills (see Fig. 2.8(a)) [79]. During the experiment, Infanoid alternates between an eye-contact behavior and a joint attention behavior. According to Kozima, the interactions of most of the children with Infanoid consist of three phases. In the first phase, children show embarrassment because they do not know how to deal with the robot. In the next phase, children gradually understand that the robot produces predictable behaviors and actively explore the capabilities of the robot. In the third, children become comfortable with the robot and treat the robot as a social agent and even start verbal conversation with it.

Keepon (see Fig. 2.8(b)) is another robot designed by Kozima and his team. It does not have a human-like shape. However, the author feels obliged to give a description of it here because of its unique design and the responses it elicits from children. Keepon is deliberately designed to have a simplistic appearance. It has a spherical head connected to a slightly larger spherical body. The head has just two eyes (microcameras) and a nose (microphone). The body contains mechanical mechanisms that allow Keepon to change its gaze direction, exhibit head nodding and shaking, and display side-to-side rocking and up-and-down bobbing behaviors. The surface of Keepon is made of silicone rubber. The robot is installed on top of a moving platform that contains a battery and wireless communication devices. Its simplistic and self-contained design makes it virtually impossible to inadvertently hurt children, so it can be safely put into a playroom where children can approach and interact with it. Longitudinal observations have been carried on how children with developmental disorders develop a bond to Keepon. Major findings from these observations are (1) that children that are considered to have underdeveloped social skills approach Keepon with curiosity and security, (2) that some of children are able to engage in diadic or triadic interactions with Keepon and (3) that each child has his/her own style during the interactions with Keepon. More detailed accounts on such interactions can be found in [81] and [80].

Simulating social development on robots

It has been shown that equipping humanoid robots with basic social skills can be achieved by careful engineering. However, infants are not born with these skills; these skills develop over time. Developmental psychologists have not yet been able to reach a consensus regarding how infants develop social skills. But it is clear that a few basic sensorimotor skills on the infant side, such as moving the head around and finding salient objects in the environment, combined with cooperation from caregivers undoubtedly play important roles over the course of infants' social skill development.

In [107], Nagai describes two models that allow a humanoid robot to develop joint attention — a basic, yet important social skill. The reasonable assumptions used by these two models are based on the observations from developmental psychology. In the context of Nagai's work, joint attention is defined as a mapping from an image that describes the head orientation of the caregiver to the motor commands that move the object the caregiver is looking at to the center of the robot's vision field. The hypothesis of a three-stage development of joint attention proposed by Buttorworth and Jarrett [28] is taken by Nagai as a ground truth.

Nagai's first model assumes that the visual acuity of the robot matures gradually so that the resolution of the images the robot receives is rather low initially but increases over time. Another assumption is the caregiver gives the robot explicit evaluation for each training sample. The evaluation criterion is initially loose and becomes more stringent as the robot's performance improves. The robot uses a four-layer neural network for learning the joint attention mapping. Learning takes place by repeating the following process: The robot first looks at the caregiver and sends the image describing her head orientation to the neural network. The output of the neural network is used to change its gaze direction. The caregiver then gives an

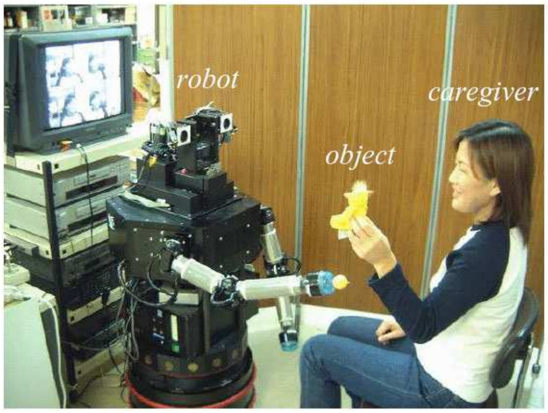


Figure 2.9: Setting used by Nagai to study the development of joint attention [107].

evaluation of the accuracy of the robot's gaze direction. This evaluation is used in turn by the robot to modify the weights of the neural network. Nagai shows that both the maturation process of the robot's visual acuity and the change of the caregiver's evaluation criterion facilitate the learning of joint attention.

In Nagai's second model, the caregiver simply changes her gaze direction to look at different objects without giving explicit evaluation of the robot's performance. In contrast to the first model where the robot has no initial skills, the robot in the second model has the innate ability of shifting its gaze to track salient objects in the environment. This innate ability is used to jump-start the learning of the joint attention mapping. Over time, the increasingly more sophisticated joint attention mapping takes over to control gaze directions. According to Nagai's results, joint attention can be successfully learned only if there are a small number of salient objects in the environment.

2.2.4 Discussion

Almost all of the research introduced above on humanoid robots are more or less inspired by the insights gained from infant development studies. For example, Ijspeer and his colleagues have used imitation learning to accelerate the transfer of human movement skills to a humanoid robot. That Leonardo is made to rely heavily on social cues to communicate with humans during collaboration is another example. However, these studies are not aimed at improving the current understanding of the development process. In contrast, Metta's work and Nagai's work study how skills such as reaching and joint attention develop on robots the way they develop in humans. More importantly, Metta, Nagai and Fitzpatrick explicitly or implicitly view development as a dynamic process where the learning of a skill may depend on more elementary skills. Metta's Babybot develops reaching upon the completion of learning basic visual-motor tasks. Fitzpatrick's Cog develops the skill of object recognition through active manipulation. The joint attention skill developed by Nagai's robot is jump-started by the ability to shift gaze direction to salient objects. The work described in this thesis falls in the same category as the work of Metta, Fitzpatrick and Nagai. Whereas each of the three projects focuses on a single domain — Metta'work on motor learning, Fitzpatrick's work on perception learning and Nagai's work on social learning — the work in this thesis spans the domains of motor learning and social learning and demonstrates the benefits of skill progression.

Chapter 3

Nico - a New Humanoid Robot

Nico is a humanoid robot designed and assembled in our lab. All the physical experiments described in the following chapters are carried out on Nico. The two sections in this chapter are devoted to the description of its hardware features and its software platform respectively.

3.1 Nico's Hardware

Nico is modeled on the 50 pecentile, male, one-year old infant [134]. It has altogether 21 degrees of freedom (DOFs), 7 in the head, 2 in the torso and 6 in each arm. Its primary dimensions are shown in Fig. 3.1. Its current design allows for more DOFs to be added to its head for facial expression display. In addition, simple graspers can be attached to its wrists to enable it to pick up small and light objects. All of Nico's structural parts are designed in Solidworks, a 3D CAD program that can produce both 3D models and 2D drawings. The drawings of finalized designs were sent to an outside company for machining. After the finished parts were received they were assembled in the lab together along with some off-the-shelf components such as bearing and gears. The main challenge of the design procedure was to find a good compromise between the size of Nico and its motion range. The larger the size, the

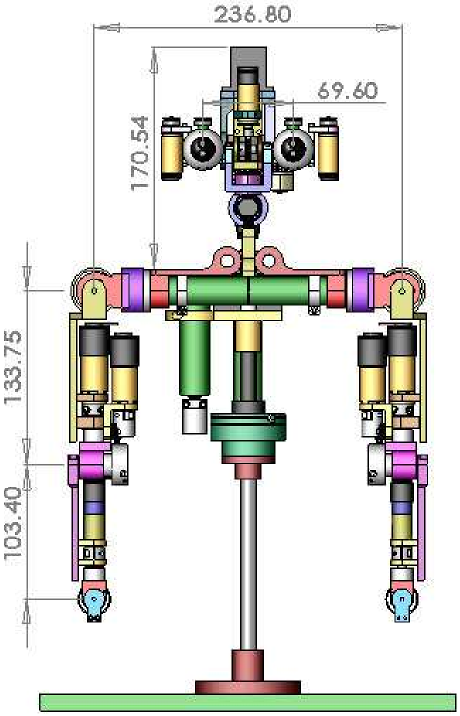
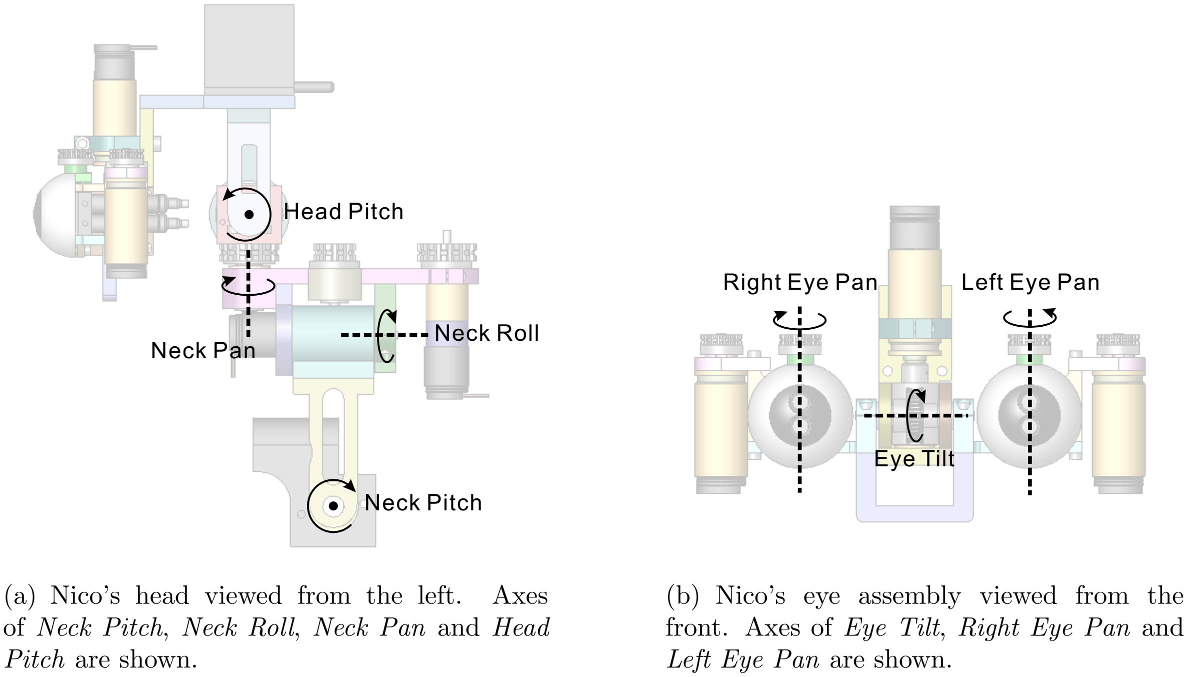


Figure 3.1: The frontal view of Nico along with its major dimensions in millimeters.

bigger the motion range. However, increasing the size too much would violate the design goal of keeping Nico about the same size as an average one-year old. During the design procedure, a considerable amount of time has been spent on tweaking the shape of various parts such that a maximum motion range is realized while the overall dimensions of Nico are kept in check.

3.1.1 The Head Design

The axes of the seven joints in the head are illustrated in Fig. 3.2. The three neck joints (*Neck Pitch*, *Neck Roll*, *Neck Pan*) together simulate the motor function of a human neck. The *Head Pitch* joint has two functions. One is to simulate head nodding without moving the whole mass of the head. The other function is to expand the range of pitch motion of the head. With *Neck Pitch* supplemented by *Head Pitch*, Nico is



(a) Nico's head viewed from the left. Axes of *Neck Pitch*, *Neck Roll*, *Neck Pan* and *Head Pitch* are shown.

(b) Nico's eye assembly viewed from the front. Axes of *Eye Tilt*, *Right Eye Pan* and *Left Eye Pan* are shown.

Figure 3.2: Head joints

able to look either straight down at the floor or straight up at the ceiling without actuating the *Eye Tilt* joint. *Neck Pitch*, *Neck Roll* and *Head Pitch* are directly driven by three DC motors while Neck Pan is driven by a belt system (see Fig. 3.3). The belt system is driven by another DC motor and its tension can be adjusted by changing the position of an idler. Nico's head can be moved into a variety of postures by actuating the neck joints and *Head Pitch* in different ways (see Fig. 3.4).

The other three joints in the head constitute a pan-tilt camera system. The *Eye Tilt* is driven by a worm mechanism (see Fig. 3.5(a)). This mechanism not only provides another stage of speed reduction, it also makes *Eye Tilt* non-backdrivable such that *Eye Tilt* can hold to its configuration when the motor driving it is turned off. The worm is supported from beneath by a sintered bronze washer that cancels out the downward thrust force exerted by the camera system on the worm and additionally provides some lubrication. The *Left Eye Pan* and *Right Eye Pan* are driven by

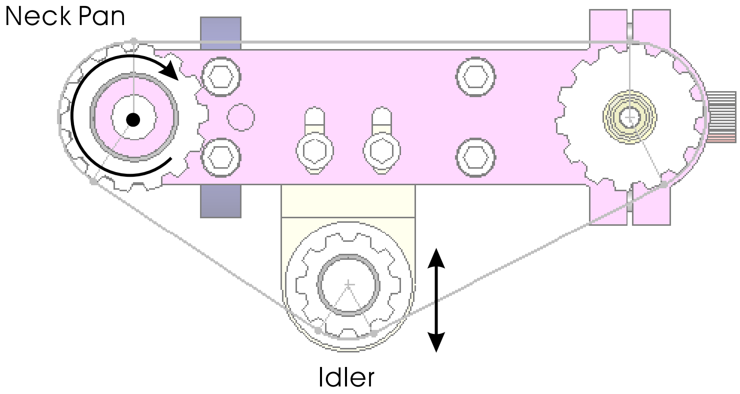


Figure 3.3: Belt system used in the *Neck Pan* joint (viewed from the top).

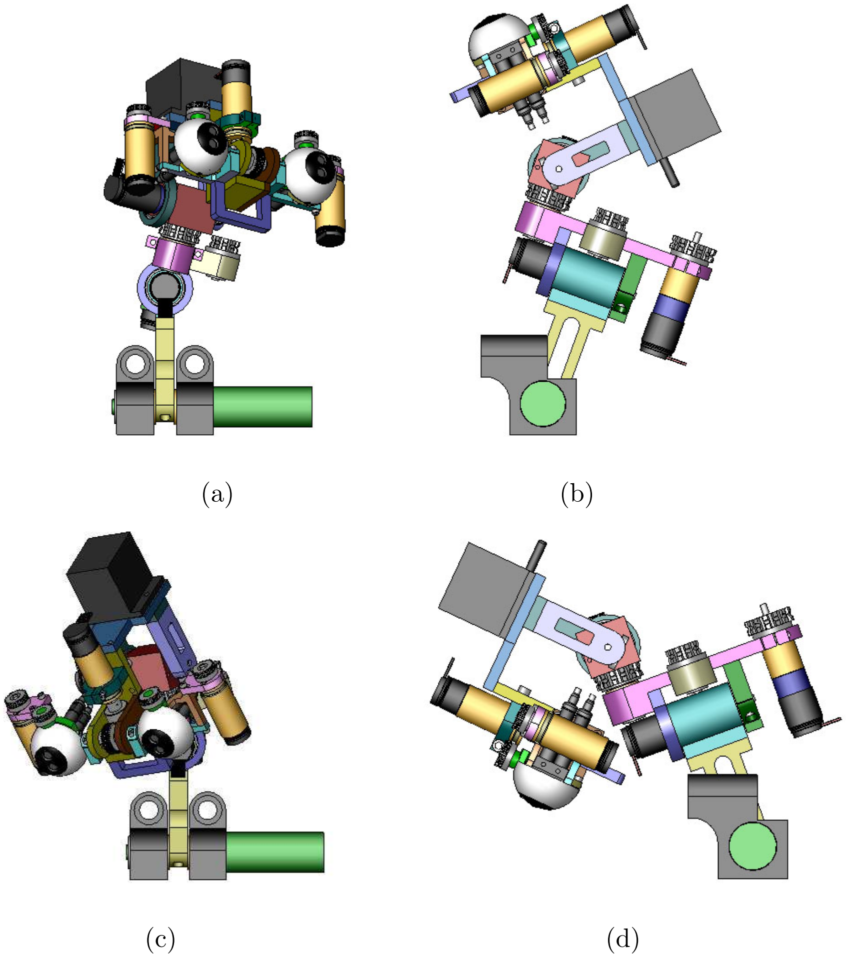
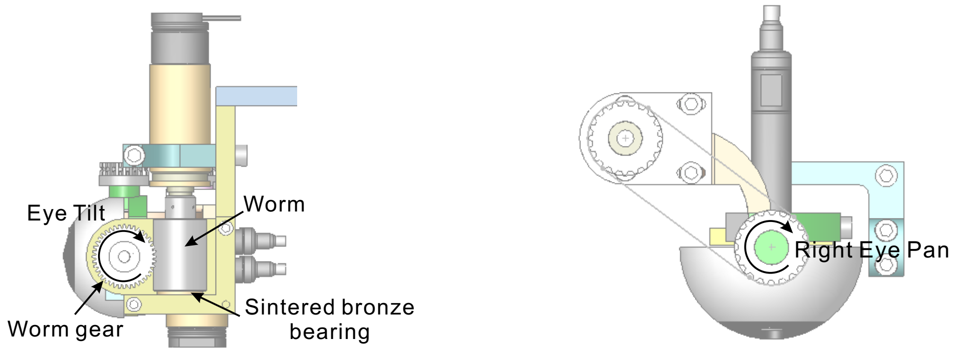


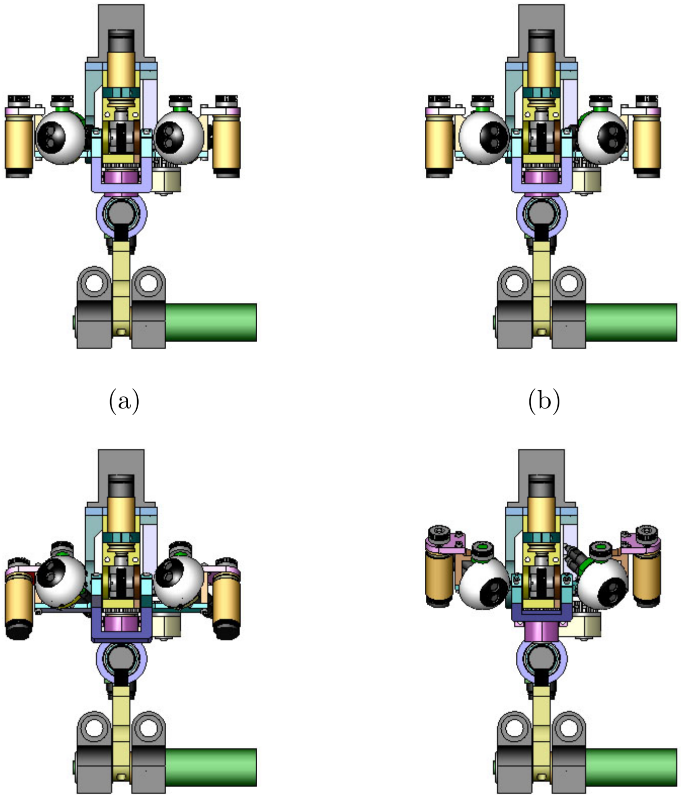
Figure 3.4: Nico can obtain a variety of head postures by actuating the three neck joints and the *Head Pitch* joint.



(a) Worm mechanism for the *Eye Tilt* joint (viewed from the left).

(b) Belt drive for the *Right Eye Pan* joint (viewed from the top).

Figure 3.5: Mechanisms in Nico's pan-tilt camera system



(a)

(b)

(c)

(d)

Figure 3.6: Pictures illustrating the range of motion of Nico's camera system

belt systems similar to the one used for *Neck Pan* (see Fig. 3.5(b)). There are two miniature CCD cameras for each eye. The long focal length camera is used for fovial vision and the short focal length camera is used for peripheral vision. A plastic "eyeball" encloses the two cameras on each side and gives Nico a more human-like appearance. The range of motion of the pan-tilt camera system is illustrated in Fig. 3.6. The pan-tilt camera system enables Nico to make saccades and track moving objects. In addition, the sensory information provided by the gyroscope can be used for simulating the vestibular-ocular reflex (VOR).

3.1.2 The Arm Design

Each of Nico's arms has six joints whose axes are illustrated in Fig. 3.7. The *Shoulder Roll* and *Shoulder Pitch* joints need to drive the whole arm, so they are driven by the most powerful motors used in Nico. Each arm is mounted onto the shoulder through a double row angular contact bearing. The shaft of the motor driving the *Lower Arm Twist* joint is extended by a custom designed shaft adapter. This adapter is squeezed onto the motor shaft by an off-the-shelf clamp. This clamp is sandwiched between two mounting pieces that help to transfer the weight of the lower arm onto the frame of the upper arm. A sintered bronze washer on top of the clamp provides some lubrication and a spring washer beneath the clamp cuts out the remaining slack. The *Wrist Twist* joint has a very similar design.

Cable drive systems are becoming popular in robotic systems. The main advantages of cable drives over traditional gear transmissions include zero backlash and design flexibility. Cable drives are used on Nico for driving the *Elbow Pitch* joint and the 1-DOF waist. In the *Elbow Pitch* joint (adapted from the elbow design in Cog [148]), a cable drive system changes the rotation axis of the dedicated motor by 90°. This system is shown in Fig. 3.8. A small pulley with an external thread and

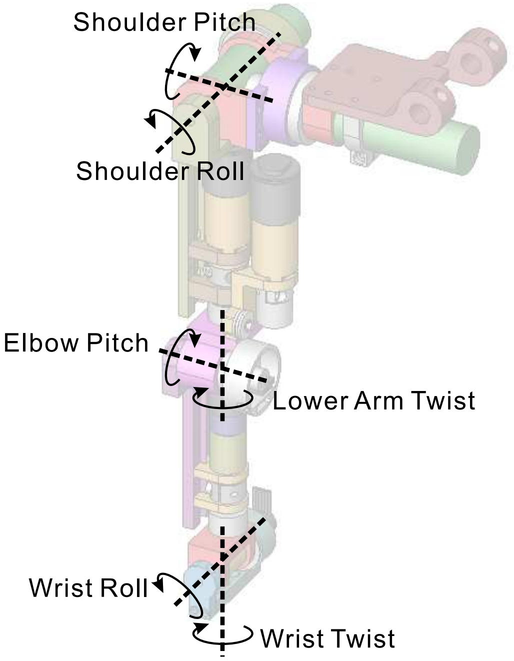


Figure 3.7: Axes of the 6 joints in the right arm (trimetric view).

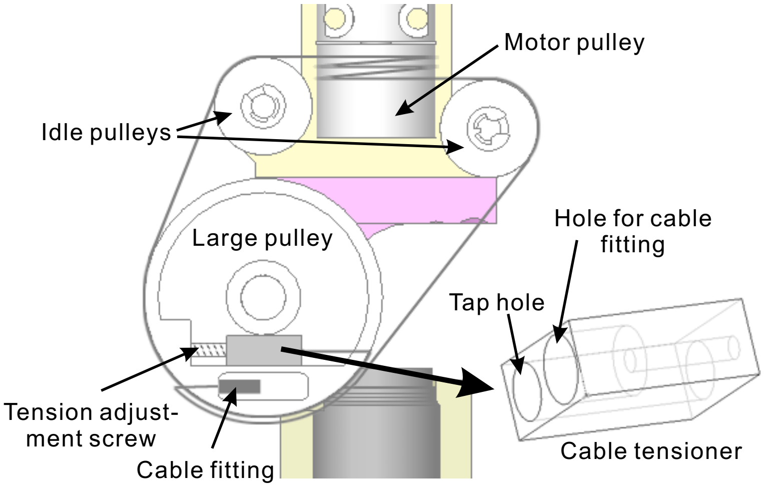


Figure 3.8: Cable drive for the *Elbow Pitch* joint.

a large pulley are attached to the motor shaft and the elbow shaft respectively. The small pulley drives the large pulley through a steel cable that is routed by two idle pulleys. In order for this system to work without backlash, proper cable tension must be maintained. Cable tension can be fine-tuned by an adjustment screw coupled with a cable tensioner. When the cable tensioner is moved to the left or to the right, the cable tension increases or decreases accordingly.

3.1.3 Additional Design Features

There is one additional joint in Nico's shoulder that gives the robot the ability to perform a shrug-like movement. It needs to be both compact and capable of delivering a large amount of torque. The final solution is shown in Fig. 3.9. The joint is powered by a DC motor that drives a worm. The worm in turn distributes the torque to the two shoulder pieces through two worm gears. This mechanism is non-backdrivable just like the worm mechanism used in the *Eye Tilt* joint. The worm and the worm gear together provide a speed reduction ratio of 20:1. This reduces the torque requirement on the driving motor. Since the worm gears exert an upward thrust force on the worm when the shoulder pieces are lifted up, a clamp and a sintered bronze washer are attached to the motor shaft to transfer the thrust force to the body.

Although all of the motors used on Nico have built-in optical encoders, they can provide a consistent reading of the motor shaft position only after they have been calibrated. For calibration, a reference point needs to be set for each motor. A simple design is used for this purpose. An example of it is shown in Fig. 3.10. In this figure, a 3D view of the *Neck Pitch* joint is shown. In the piece that supports the shaft of the driving motor, a small hole is drilled. A short stainless steel shaft is pushed into this hole and used as a limiter. During calibration, the lower neck piece is turned slowly in the clockwise direction until it hits the limiter and stalls. The position of the

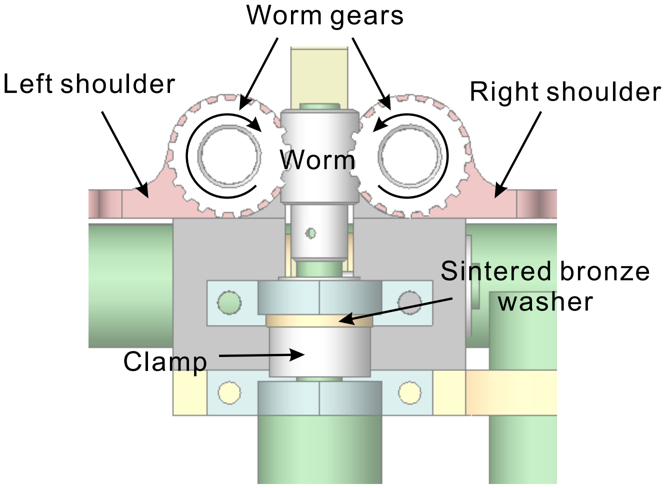


Figure 3.9: Mechanism that gives Nico the ability to shrug (view from the back).

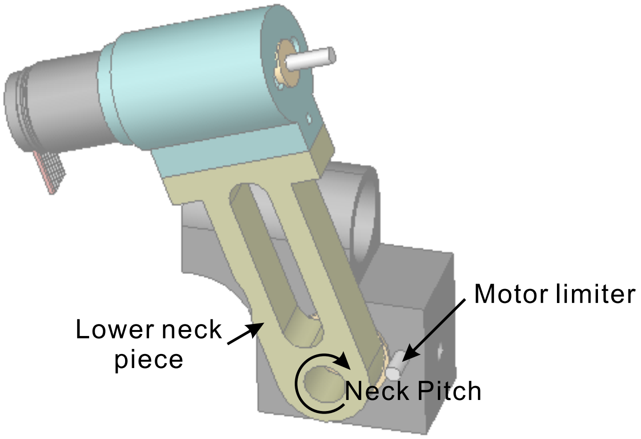


Figure 3.10: A simple mechanism for motor calibration.

motor shaft at this moment is set as the zero position upon which all future encoder readings of this motor will be based.

Nico's 1-DOF waist is also implemented with a cable drive system. This system is however simpler than the one used in Elbow Pitch since no change of rotation axis is necessary. In contrast to Elbow Pitch, the large pulley in the waist is kept stationary and the small pulley along with the upper body revolves around it.

3.1.4 Motor Selection

Motor selection is a critical part of the whole design procedure. Motors constitute about one third of the total mass of Nico. Because of the constraints put on the dimensions of Nico, the sizes of the motors need to be kept as small as possible. On the other hand, the motors on Nico must deliver enough power and torque to make the robot as agile as possible. Micromo and Maxon are the two leading manufacturers for miniature motors. Both companies can deliver parts in small quantities within a very short period of time. Therefore, only the products from these two companies are considered for the design of Nico. All of the motors we purchased for Nico are preassembled with a gear head for speed reduction and a high-resolution optical encoder for keeping track of the shaft position.

The first step of motor selection is to estimate the maximum torque T_{spec} and the required speed S_{spec} at this torque for a particular joint. This estimation should take in consideration factors such as whether there is an extra speed-reduction stage in the joint. After T_{spec} and S_{spec} have been estimated, for each candidate motor+gearhead assembly, the torque T_m the motor must produce to make the output torque of the assembly be T_{spec} is calculated by dividing T_{spec} by the product of the ratio and the efficiency of the gearhead. T_m should be no larger than half of the motor's stall torque to ensure good service life. The speed output S_m of the motor at T_m is calculated by subtracting the product of T_m and the slope of the speed-torque curve of the motor from the no-load speed of the motor. Dividing S_m by the ratio of the gearhead results in the output speed S_o of the motor+gearhead assembly. S_o should be no smaller than S_{spec}. After such evaluations, the selection of a specific assembly from all the qualified ones is determined by further comparison of features such as length, weight and the amount of backlash at the output.

3.2 Nico's Software

All sensors and motors on Nico are connected to a 16-node computation cluster running the QNX real-time operating system. Nodes are connected through a gigabit backbone switch and a number of point-to-point network links. Two additional computers running Windows and Linux respectively are also connected to the backbone switch. These machines run softwares such as Small Vision System (for stereo vision) and Watson (for head pose estimation) that are currently not available on QNX.

The software package of Nico consists of a large number of small modules. Each of them implements a specific function. Many of these modules are low-level drivers that communicate directly with the sensors and motors on Nico. Other modules implement higher-level functions such as visual attention and object tracking. The internal workings of individual modules will be described in detail in the following chapters.

During run-time, selected modules are allocated to processing nodes based on their computation requirements. Active modules can selectively communicate with one another through a special communication interface. This communication interface is based on two main classes - TcpIpSender and TcpIpReceiver. To send and receive data, a module needs to instantiate a sender object based on TcpIpSender and an received object based on TcpIpReceiver respectively. A sender object can be connected to multiple receiver objects so that one data piece can be propagated to multiple modules with only one function call at the application level. Within each connection, multiple channels can be established, each handling a unique type of data. The times on all nodes in the computation cluster and the two external computers are synchronized using Network Time Protocol (NTP). A module can examine the "freshness" of the data pieces it receives by checking their embedded time stamps. In this way, mistakes caused by mixing up information originated at different times can be avoided.

Chapter 4

Learning to Reach

In this chapter, a model for learning reaching movements toward spatial targets is presented. It consists of two main components. The first component is a learned model of the forward kinematics of the arm (often called the forward model of the arm), which transforms joint angles of the arm into the position of the arm end effector. The second component is a directional mapping that maps a desired movement direction of the end effector into angular increments of the arm joints. It is derived from the forward model and does not need to be learned separately. Based on the information provided by these two components and the visual perception of the arm end effector (whenever available), accurate reaching movements toward spatial targets can be generated. Compared with other published methods, the model proposed in this chapter requires far fewer training samples and a smaller representation space. It is also consistent with psychological and physiological observations. Its performance is demonstrated through both simulations and physical experiments on Nico.

The first section of this chapter reviews some of the previous work on reaching. The next section describes our new model in detail, putting the emphasis on how the forward model is learned and how reaching movements are generated. Following that, the physical experiments carried out on Nico to test the model in the real world are presented. This chapter is concluded with a final discussion.

4.1 Related Works

In robotics, early work on reaching focused primarily on inverse kinematics. Many of the solutions were based on the Resolved Motion Rate Control (RMRC) algorithm which requires the forward kinematics of the arm to be known for the computation of Jacobian matrices [146]. In contrast to high-precision robotic manipulators, the arm kinematics of a human is difficult to describe as a closed-form function and changes over an individual's lifespan, especially in the early years. In humans, proficiency in reaching is not present immediately after birth; it must be learned through practice over time. As introduced in Section 2.1.5, infants typically make their first successful reach at the age of three to five months. They can reliably reach spatial targets about three months later. However, adult-level proficiency is only achieved after many more months of practice.

What must be learned in order to achieve accurate reaching movements if the values of the arm parameters are not readily available to the learning system? One seemingly straightforward approach is to learn the inverse kinematic mapping from x_{target} to θ directly, where x_{target} denotes the task vector that describes the target position and θ denotes the joint vector describing the corresponding arm posture. This approach is problematic in the case of human arm movements and many humanoid robot arm movements because the dimension of the joint vector is larger than that of the task vector. The result of this redundancy is that there can exist multiple values of θ which correspond to the same x_{target}. Choosing an appropriate one from all possible values of θ can be difficult [39].

Jordan and Rumelhart suggested an approach that circumvented the difficulty mentioned above [74]. Instead of learning the inverse kinematics of a redundant arm directly, a model of the forward kinematics of the arm is learned with a neural network. Such a model is often called *forward model*. (There is physiological evidence that forward models are used by the brain for a variety of sensorimotor tasks [94].) After learning the forward model, a second neural network is constructed and its output neurons are directly connected to the input neurons of the network representing the forward model. The combined network accepts task vector x as input and outputs another task vector x'. In the next step, the connective weights in the second network are modified through error backpropagation such that the combined network represents an identity function in the end. At this point the second network is disconnected from the forward model and is used afterwards for computing a joint vector θ for any given target vector x_{target}.

While Jordan and Rumelhart's approach yields a one-to-one inverse mapping of the arm, this inverse mapping remains fixed after learning. This, in effect, removes the flexibility brought by the redundancy in the arm. Moreover, their approach does not provide a measure to reduce the residual errors in the inverse mapping. Bullock et al. suggested that a better solution to the reaching problem should contain both a learned forward model of the arm and a learned mapping that maps a direction vector in the task space to a direction vector in the joint space. Instead being used to obtain a unique inverse mapping of the arm, the forward model in Bullock et al.'s solution serves the purpose of predicting the current position of the arm end effector. Arguments supporting the learning of both a forward model and a directional mapping have also been put forth by other researchers [93][125][126]. In particular, Miall compared existing solutions to the reaching problem from the perspective of control engineering [93]. From this perspective, a directional mapping acts as a feedback controller and a forward model stabilizes the system by removing the sensory delays.

A variety of learning algorithms have been suggested for acquiring these functions, including Self Organizing Maps (SOM) [119][25], Multiple Layer Perceptrons (MLP) [74] and Locally Weighted Projection Regression (LWPR) [41]. The choice of learning algorithm determines directly the compactness of the representations of the forward model and the directional map, as well as the number of training samples required to ensure good convergence. The larger the dimension of the joint vector θ, the greater the need for compact representations and reduced training set sizes. Bernstein identified this as the "degrees of freedom problem" [11] and it is known in statistical learning theory as the curse of dimensionality [66].

While many have argued about which mappings should be learned and which learning algorithms are most appropriate for this task, the dimensionality problem remains. As an example of scale, Bullock et al. employed 15625 neurons to learn a forward model of a 3-DOF arm based on a 40000-sample training set [25]. Besides the dimensionality problem, the question of how perceptual noise affects the accuracy of reaching has not been sufficiently addressed in the literature. Also, the performance of the previously proposed models has been demonstrated primarily in simulations; evidence that they also work in the real world is lacking.

4.2 A New Model for Learning to Reach

Fig. 4.1 provides an overview of a novel model for learning to reach. The basis of the model is a forward model of the arm that is learned autonomously through motor babbling. It is important that the parameters for learning the forward model are set appropriately in order to keep the required number of training samples and the size of the representation space as low as possible. After the forward model is learned, it is used during a reaching movement to predict the position of the end effector from the current arm posture. The pseudo-inverse of the Jacobian corresponding to the current arm posture is computed directly from the forward model. Δx is a direction vector of small magnitude in the task space pointing from the predicted end effector position to the target position. It is transformed into a joint vector increment $\Delta\theta$ that is relayed to the arm motors. If the end effector can be perceived by the stereo vision system during the reaching movement, x_{pred} is replaced by x_{perc} and Δx becomes a direction vector pointing from the perceived end effector position to the target position. A description of the stereo vision component is given in Section 4.2.1. The training of the forward model and the algorithm for incremental trajectory generation are described in detail in Sections 4.2.2 and 4.2.3 respectively.

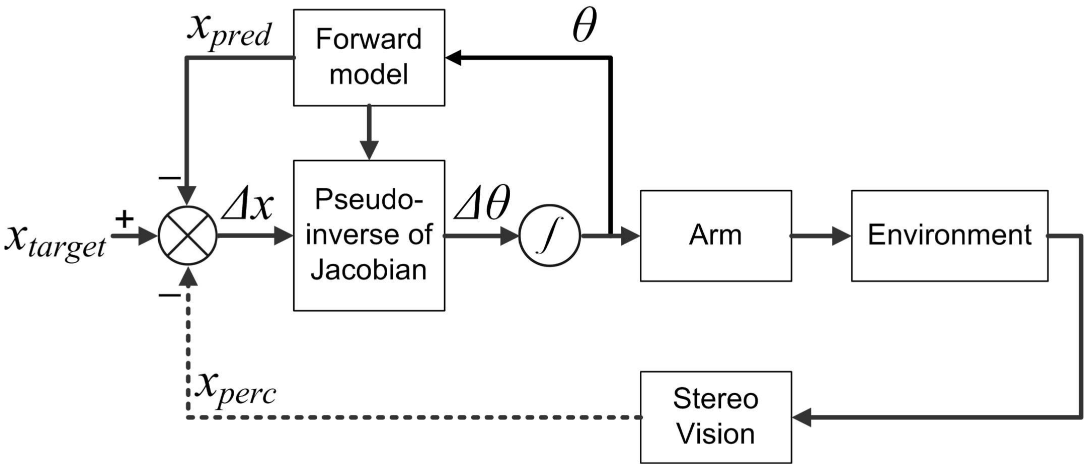


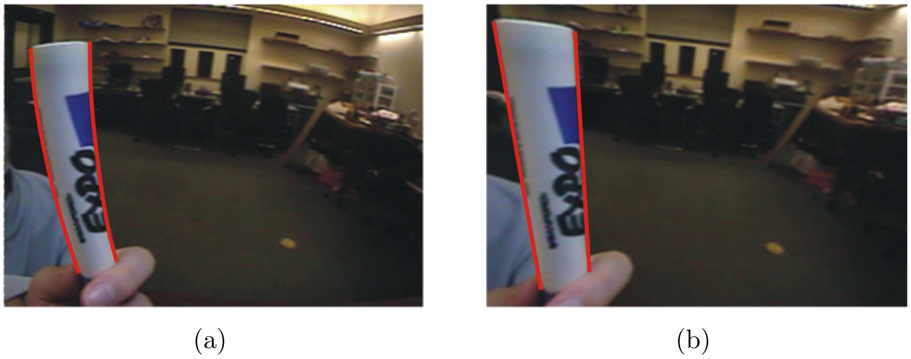
Figure 4.1: Overview of a new model for learning to reach based on a learned forward model of the arm kinematics. During a reaching movement, the forward model is used both for the prediction of the end effector position and the computation of the pseudo-inverse of the current Jacobian. The dashed line representing the data flow from the stereo vision component indicates that visual feedback is optional during reaching.

4.2.1 Stereo vision

The stereo vision component shown in Fig. 4.1 retrieves video data from the two short focal length cameras as input. The two long focal length cameras prove to be impractical for the stereo vision needed for the reaching behavior because their common vision field has only a small overlap with the reachable space of the robot arm. The radial distortion coefficients K_1 and K_2 and other parameters are measured for each camera through the Camera Calibration Toolbox for Matlab developed by J.-Y. Bouguet [18]. The pixel value of position (x, y) in the corrected frame is filled with the pixel value of position (x', y') in the original frame through the following equation [69]:

$$\begin{aligned} x' &= x(1 + K_1 r^2 + K_2 r^4) \\ y' &= y(1 + K_1 r^2 + K_2 r^4) \end{aligned} \tag{4.1}$$

If values of x' and y' are not integers, they are replaced by their respective integer parts. During the startup of the vision subsystem, a lookup table is built for each camera consisting of the mappings from (x, y) to (x', y') for all possible (x, y). The pre-built lookup tables enable an efficient distortion correction in real time.



(a)

(b)

Figure 4.2: Original image (a) and image (b) corrected for radial distortion through Eq. (4.1)

Currently, a ϕ19.05mm wooden ball is attached to the distal end of the robot arm as the end effector and a wooden ball of the same diameter is used as the reaching target. The projection of either the end effector or the target on each camera image plane is replaced by its centroid. For the experiments described in this chapter , the two eye cameras are positioned parallel to each other. These simplifications make it straightforward to determine the position of both the end effector and the target in the eye-centered coordinate system [127].

4.2.2 Learning a forward model

The forward kinematic function of the arm is defined as a mapping $f_{arm} : \theta \rightarrow x$. Each of Nico's arms has 6 degrees of freedom, so we have $\theta \in R^6$. At the current stage, we require our robot to touch a presented target without putting any restriction on the orientation of its end effector. This makes $x \in R^3$. A forward model of the arm is learned through motor babbling, during which the arm is repeatedly moved into random postures. If the end effector can be perceived by the stereo vision system at the end of an arm movement, the joint vector θ corresponding to the current arm posture and the perceived 3D position x of the end effector is recorded as a training sample. The arm stops moving only when a pre-specified number of training samples are gathered.

Learning a forward model of the arm is essentially approximating the function f_{arm} through training samples of the form $(\theta_i, x_i)_{i=1,2,...,n}$, where n is the size of the training set. Each x_i in (θ_i, x_i) contains noise introduced by the stereo vision system. Neural networks such as MLP and RBFN are commonly used function approximation techniques [68]. The most important reason for our adoption of RBFN for learning the forward model is that the only weights to be learned are those connecting the hidden layer and the output layer. They can be determined directly by the linear least square method, which avoids the problem of local minima that MLPs often encounter. Arguments favoring RBFN from the perspective of biological plausibility will be given in Section 4.2.2.

Optimization of learning parameters

The Gaussian function is used as the basis function in the hidden layer of our RBFN. It can be expressed as $g(x) = exp(\|x-c\|\cdot 0.8326/spread)$, where x is the input vector and c is the center of the Gaussian. The parameter *spread* controls the extent of g's influence in its neighborhood. Since it has been shown that a RBFN with a set of basis functions that have a common form but different centers can approximate any continuous input-output mapping [112], the same value is assigned to the *spread* of each Gaussian. We use the Orthogonal Least Squares (OLS) algorithm to determine the number and the centers of the Gaussian functions automatically from the training data [34]. The training stops when the root mean square error ($rmse$) of the network falls below a certain *margin*. The optimal values for both *spread* and *margin* must be determined before training. In addition, we are also interested in keeping the size of the training set as small as possible to save the time spent on gathering training samples. Computer simulations are used to optimize the three learning parameters (*spread*, *margin* and training set size) because it is very difficult to measure accurately the quality of the learned forward model through physical experiments.

Since the optimal *spread* of the Gaussian functions in the hidden layer depends primarily on the actual function to be approximated, the effect of noise is excluded in the simulations conducted to optimize *spread*, which means the value of x_i in the training sample (θ_i, x_i) is for the time being the true end-effector position corresponding to the joint vector θ_i. Fig. 4.3 shows four *rmse*-*spread* curves for four different training set sizes. The line style of each curve indicates which size it corresponds to. Each data point is generated by averaging the *rmse* values measured on 40 random

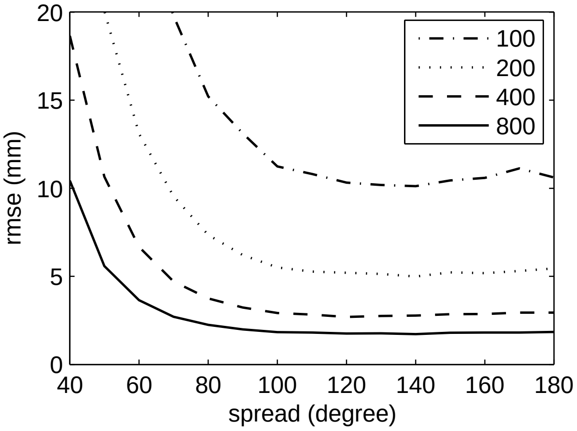


Figure 4.3: *rmse-spread* curves for four different training set sizes. The averaged optimal *spread* is 130. (Data generated by simulations.)

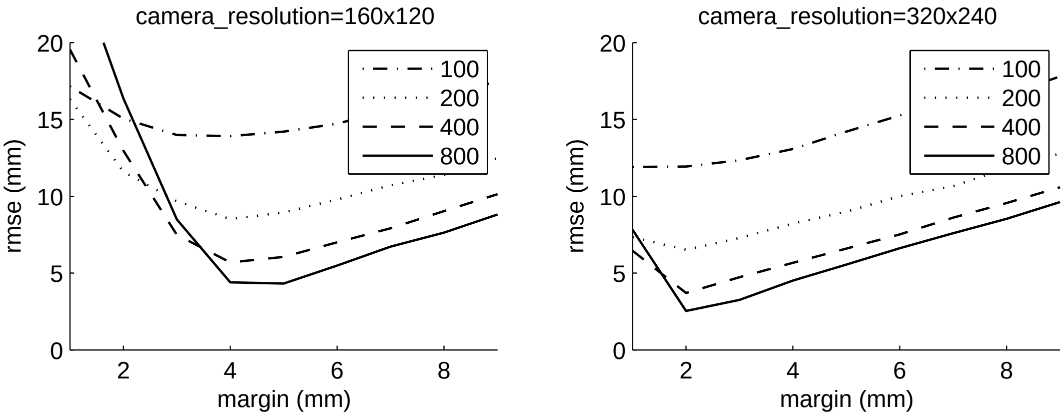


Figure 4.4: *rmse-margin* curves for four different training set sizes and two camera resolutions. For each camera resolution, the optimal *margin*s for different training set sizes lie quite close to each other. (Data generated by simulations.)

training sets of the same size. It can be observed that for all of the four curves displayed, the *rmse* falls sharply at the beginning. As *spread* grows larger, the *rmse* appears to stay constant although it actually rebounds very gently after reaching a minimum point. The average of the optimal *spread* values of the four *rmse-spread* curves is about 130.

A large part of the noise in the perceived 3D position of the end effector is the stereo perception error, which is dependent on the resolution of the camera images. To find the optimal values of *margin* for different camera resolutions, simulations were conducted that use training samples whose task vector components contain stereo perception errors. *spread* was set to a constant value of 130. The results of the simulations to optimize *margin* are shown in Fig. 4.4, where the curves are plotted in different styles according to the same convention used in Fig. 4.3. As expected, the optimal *margin* for a 160x120 camera resolution is more than 2mm higher than that for a 320x240 camera resolution. Lower-resolution images result in higher stereo perception error which in turn requires a larger value for *margin* to prevent overfitting. Fig. 4.4 also shows that the more training samples we use, the higher the quality of the learned forward model. However, a 400-sample training set already leads to a small *rmse* very close to that achieved by a 800-sample training set. 400 samples require a very moderate amount of time to gather on a physical robotic platform, less than 30 minutes in our case.

In the real world, the stereo perception error only partially contributes to the noise in the perceived position of the end effector. The end effector of a robotic arm is a 3D structure. Its projection on the camera image plane is not a point, but a blob. Using the centroid of the blob as we do for the calculation of the end effector position introduces additional error since the two centroids on the right and left image planes do not represent the same point on the end effector. Putting a special marking on the end effector does not solve the problem effectively because it is hard to guarantee that this marking is visible for all arm postures. From Fig. 4.4, we can see that *rmse* rises significantly as *margin* moves away from its optimal value. For a RBFN learned through motor babbling on a physical robot, it is difficult, if not impossible, to

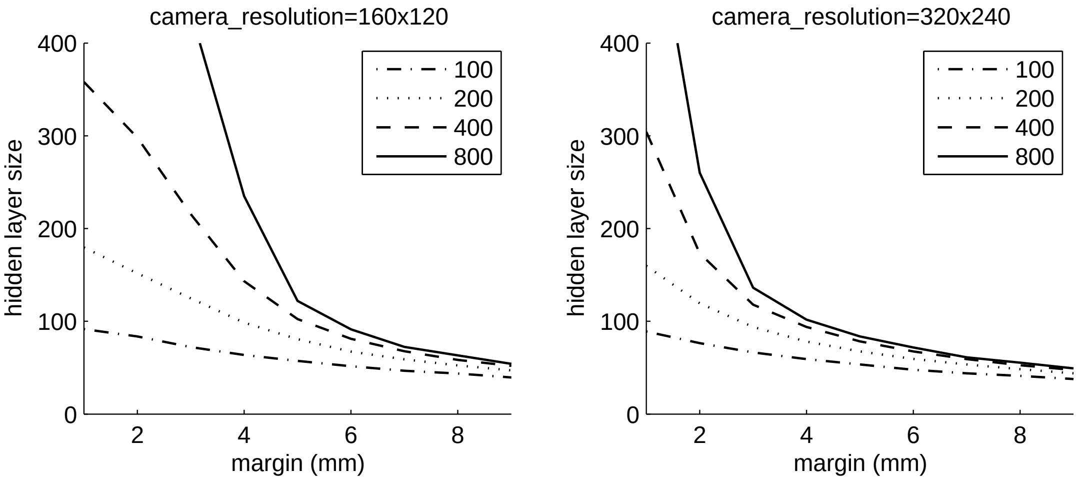
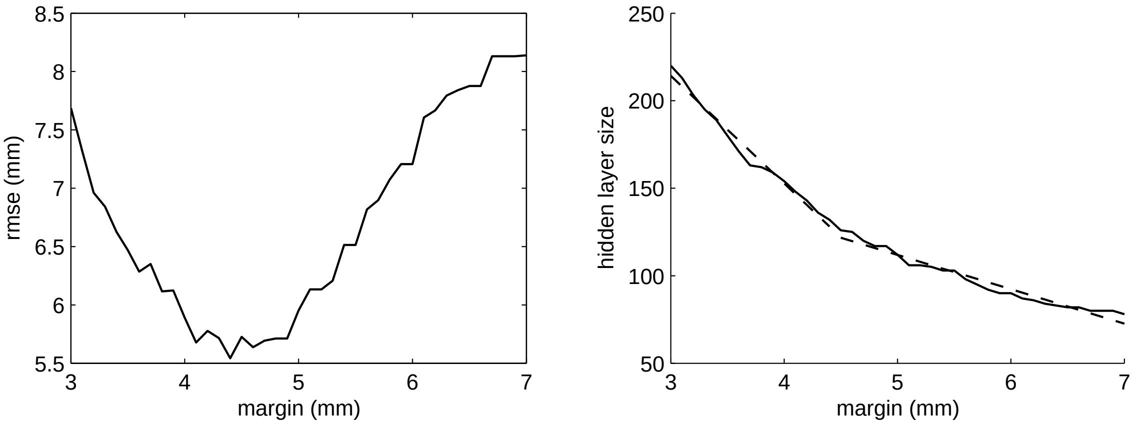


Figure 4.5: Hidden layer size versus *margin* curves for four different training set sizes and two camera resolutions. Unlike the *rmse-margin* curves that are difficult to generate on a physical robotic platform, these curves can be easily produced. (Data generated by simulations.)



(a)

(b)

Figure 4.6: (a) shows the *rmse-margin* curve of a RBFN trained on a 400-sample set. (b) shows that the hidden layer size versus *margin* curve of the same network can be fitted very well with two straight lines (dashed). The intersection of these two lines corresponds roughly to the optimal *margin*. A 160x120 camera resolution is used during motor babbling. (Data generated by simulations.)

measure the true error vector $f_{arm}(\theta_i) - \tilde{f}_{arm}(\theta_i)$ accurately, where $\tilde{f}_{arm}$ is the learned forward model represented by a RBFN. Thus, finding out the optimal *margin* on a physical robot through gathering data to plot the *rmse-margin* curve is impractical.

Fortunately, the representation of the RBFN provides a measurement that is easy to gather and useful for determining the optimal *margin*. Fig. 4.5 shows the curves of the hidden layer size of the RBFN versus *margin*. Many of those curves appear to be consisted of two straight segments whose intersection corresponds approximately to the optimal *margin*. Some curves exhibit a third segment in the middle, but this phenomenon is caused by the coarse increment for *margin* used for generating these curves. This observation leads to the following practical strategy: First generate a curve of hidden layer size versus *margin* by training a new RBFN each time with a slightly larger *margin* on the same training set. Then find the splitting point e for which the sum of the errors for fitting the left and the right part of the curve with two different straight lines is minimal. e can be found with a simple exhaustive search. Fig. 4.6 shows an example.

A biologically plausible implementation

Radial basis function networks have a solid theoretic foundation and close ties to regularization theory and Support Vector Machine (SVM) [117][113][59][138]. Poggio has suggested that RBFN is one of the learning mechanisms in the brain [114]. Pouget et al. saw RBFNs as a form of population coding that could play an important role in sensorimotor transformation [115][40].

Our implementation of RBFN uses the OLS algorithm to determine the number and the centers of the basis functions in the hidden layer. OLS is a sophisticated algorithm that tries to approximate an unknown function with a minimum number of hidden neurons; it is highly unlikely to be used by a biological learning system. However, OLS is not essential for RBFN training; other algorithms or even heuristics can be used as substitutes. To consider the worst-case performance, we compared the performance of two RBFNs, one of which is trained with OLS while for the other, the

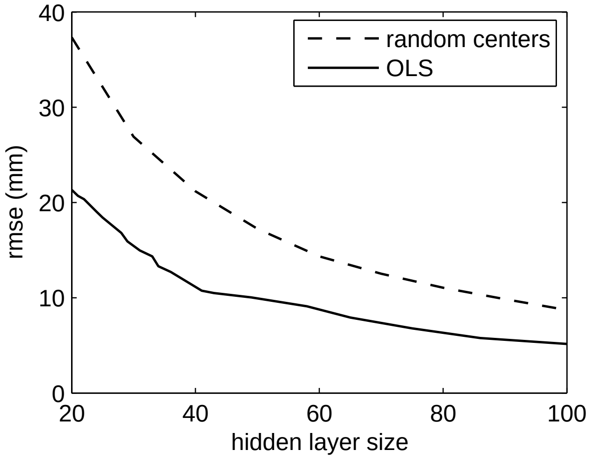


Figure 4.7: The solid curve is generated by training a RBFN repeatedly on a 400-sample set with OLS, each time with slightly larger *margin*. Each data point on the dashed curve is the averaged *rmse* of a RBFN trained on 40 different 400-sample sets. Each time, the number of the basis functions in the RBFN is kept the same, but their centers are randomly set to the joint vectors of randomly selected samples from the training set. A 320x240 camera resolution is used during motor babbling. (Data generated by simulations.)

centers of the hidden layers are set to the joint vectors of randomly selected samples from the training set. The results of the simulation are shown in Fig. 4.7. It can be observed that the *rmse* of the second network is within a factor of 2 of that of the first network if both networks use the same number of hidden neurons. Compared with the overall range of motion of the arm, this error is still quite small.

Once the number and the centers of the basis functions in the hidden layer have been determined, the weights from the hidden layer to the output layer can be calculated directly. The output of the hidden layer for the task vector θ_i of the training sample (θ_i, x_i) can be described with vector h_i. Using $H = [h_1\ h_2\ ...]$ and $X = [x_1\ x_2\ ...]$ respectively, we would like to find a weight matrix W such that

$$WH = X. \tag{4.2}$$

Usually no exact solution can be found. A minimum norm solution can be computed through the linear least square algorithm as

$$W = XH^T(HH^T)^{-1}. \tag{4.3}$$

From another perspective, W in Eq. (4.2) can be seen as an associative memory. It can be formed through an incremental learning rule such as the biologically plausible Hebbian learning, which is the only learning mechanism needed to learn a forward model of the arm with a RBFN since the centers of the basis functions in the RBFN can be randomly selected as discussed in the previous paragraph [119][68].

4.2.3 Incremental trajectory generation

Algorithm description

The forward kinematics equation $x = f_{arm}(\theta)$ can be transformed into

$$\dot{x} = J(\theta)\dot{\theta} \tag{4.4}$$

by taking derivative on both sides. Resolved Motion Rate Control (RMRC) solves Eq. (4.4) with

$$\dot{\theta} = J^{\#}\dot{x}, \tag{4.5}$$

where $J^{\#}$ is the pseudo-inverse of the Jacobian matrix J [146]. $\dot{\theta}$ in Eq. (4.5) is the minimum norm solution to $\dot{x}$ satisfying Eq. (4.4). Liegois proposed an extension

$$\dot{\theta} = J^{\#}\dot{x} + \alpha(J^{\#}J - I_n)\nabla H \tag{4.6}$$

that exploits the null space of J to incorporate an additional optimization criterion H into Eq. (4.5) [84]. Typical applications of Eq. (4.6) include singularity avoidance and obstacle avoidance [108][87]. We use the original form of RMRC as the basis for our incremental trajectory generation algorithm (ITGA). However, it should be noted that our ITGA can be easily adapted to any extension of RMRC based on Eq. (4.6).

Eq. (4.5) can be approximated by

$$\Delta\theta = J^{\#}\Delta x. \tag{4.7}$$

$J^{\#}$ in Eq. (4.7) can be understood as a directional mapping which transforms the direction vector Δx in the task space into direction vector $\Delta\theta$ in the joint space. Bullock et al. suggested that an approximation of the directional mapping $J^{\#}$, denoted as $\tilde{J}^{\#}$, can be learned independent of the forward model $\tilde{f}_{arm}$. However, the most direct way to obtain $\tilde{J}^{\#}$ is to extract $\tilde{J}$ from $\tilde{f}_{arm}$ and transform it into $\tilde{J}^{\#}$. (Recall that $\tilde{f}_{arm}$ can be easily learned as was shown in Section 4.2.2) One way to extract $\tilde{J}$ is to replace the basis functions in the hidden layer of the RBFN representing $\tilde{f}_{arm}$ with their appropriate partial derivatives [16]. For instance, in order to get an approximation of $[J_{11}, J_{21}, J_{31}]^T$, we simply replace the basis functions by their partial derivatives with respect to θ_1 and use the output of the network as the approximation for $[J_{11}, J_{21}, J_{31}]^T$. In this way, a complete Jacobian approximation $\tilde{J}$ can be constructed. Another way to derive $\tilde{J}$ is simply to use numerical differentiation. Our ITGA that relies on both $\tilde{f}_{arm}$ and $\tilde{J}$ is listed in Table 4.1. This version assumes that visual feedback of the end effector position is not available throughout the reaching movement.

Based on the predicted current position of the end effector $\tilde{x}(i-1)$ and the target x_{target}, the joint vector $\tilde{\theta}(i)$ at step i is calculated using

$$\tilde{\theta}(i) = \tilde{\theta}(i-1) + \tilde{J}^{\#}\alpha(i)(x_{target} - \tilde{x}(i-1)), \tag{4.8}$$

where $x_{target} - \tilde{x}(i-1)$ represents the direction vector in the task space toward the

Table 4.1: Incremental Trajectory Generation Algorithm (ITGA).

```
**initialize** $\theta_{start}, x_{target}, step\_size$
$\tilde{\theta}(0) \leftarrow \theta_{start}$
$\tilde{x}(0) \leftarrow \tilde{f}_{arm}(\tilde{\theta}(0))$
$i \leftarrow 0$
**loop**
    $i \leftarrow i + 1$
    **if** $\|x_{target} - \tilde{x}(i-1)\|_2 > step\_size$
        $\alpha(i) \leftarrow step\_size / \|x_{target} - \tilde{x}(i-1)\|_2$
    **else**
        $\alpha(i) \leftarrow 1$
    **end**
    Calculate $\tilde{J}$ and $\tilde{J}^{\#}$ for $\tilde{\theta}(i-1)$
    $\Delta x \leftarrow \alpha(i)(x_{target} - \tilde{x}(i-1))$
    $\tilde{\theta}(i) \leftarrow \tilde{\theta}(i-1) + \tilde{J}^{\#}\Delta x$
    Output $\tilde{\theta}(i)$ to the motor controller
    $\tilde{x}(i) \leftarrow \tilde{f}_{arm}(\tilde{\theta}(i))$
    **if** $\alpha(i)$ **equal** 1
        **break**
    **end**
**end**
```

reaching target. $\alpha(i)$ ensures that the magnitude of $\alpha(i)(x_{target} - \tilde{x}(i-1))$ is equal to *step_size* before the final reaching step. Eq. (4.8) can be seen as a further approximation of Eq. (4.7).

Analysis

To make it clear that both the learned forward model $\tilde{f}_{arm}$ and the extracted Jacobian $\tilde{J}$ are involved in Eq. (4.8), we can rewrite Eq. (4.8) as

$$\tilde{\theta}(i) = \tilde{\theta}(i-1) + \tilde{J}^{\#}\alpha(i)(x_{target} - \tilde{f}_{arm}(\tilde{\theta}(i-1))). \tag{4.9}$$

Eq. (4.9) shows that the errors in both $\tilde{J}$ and $\tilde{f}_{arm}$ contribute to the final position error. In order to isolate the effects of these two kinds of errors, we consider three variants of Eq. (4.9), which rely upon exact models:

$$\tilde{\theta}(i) = \tilde{\theta}(i-1) + J^{\#}\alpha(i)(x_{target} - \tilde{f}_{arm}(\tilde{\theta}(i-1))). \tag{4.10}$$

$$\tilde{\theta}(i) = \tilde{\theta}(i-1) + \tilde{J}^{\#}\alpha(i)(x_{target} - f_{arm}(\tilde{\theta}(i-1))) \tag{4.11}$$

and

$$\tilde{\theta}(i) = \tilde{\theta}(i-1) + J^{\#}\alpha(i)(x_{target} - f_{arm}(\tilde{\theta}(i-1))) \tag{4.12}$$

In contrast to Eq. (4.9), Eq. (4.10) only uses $\tilde{f}_{arm}$ to generate a reaching trajectory while Eq. (4.11) only uses $\tilde{J}$. So the final position error resulting from an ITGA based on Eq. (4.10) or (4.11) is caused solely by the error in $\tilde{f}_{arm}$ or $\tilde{J}$ respectively. Eq. (4.12) can be seen as an exact reformulation of Eq. (4.7). Because Eq. (4.12) relies on the exact forward kinematics and Jacobian, among the four equations listed above, it should achieve the best reaching performance and can serve as a benchmark. For the sake of simplicity, the four different versions of ITGA based on Eq. (4.9), (4.10), (4.11) and (4.12) are referred to below as ITGA($\tilde{f}_{arm}, \tilde{J}$), ITGA($\tilde{f}_{arm}, J$), ITGA($f_{arm}, \tilde{J}$) and ITGA($f_{arm}, J$) respectively.

Error histograms of the four ITGAs defined above are shown in Fig. 4.8. They are generated by simulations based on the same forward model. A 320x240 camera resolution is used during motor babbling. It is assumed that x_{target} contains no perceptual noise. 10mm is assigned to the variable *step_size*. The starting posture is selected such that all joints assume values in the middle of their motion ranges. Since ITGA($\tilde{f}_{arm}, \tilde{J}$) uses no visual feedback at all, it is very satisfying to see that almost all of its errors are below 5mm. Despite the good reaching performance achieved, there are two very interesting questions to be answered: What are the most important factors that determine the reaching accuracy? Does our model scale to tasks requiring

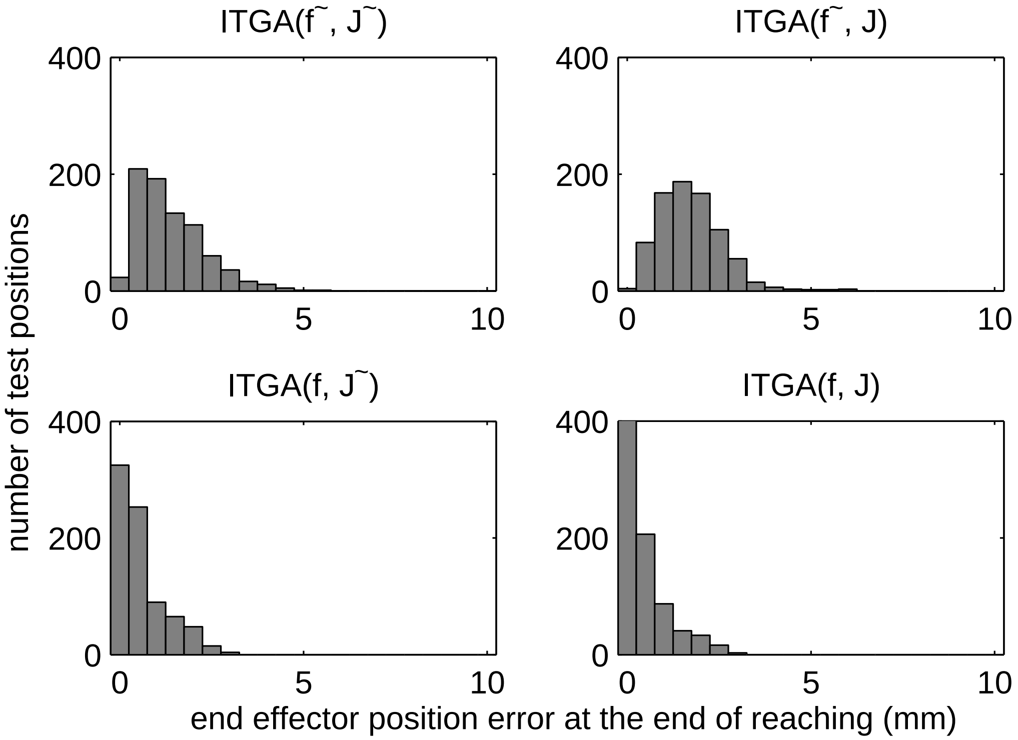


Figure 4.8: Error histograms of the four different versions of ITGA. A 320x240 camera resolution is used during motor babbling. 10mm is assigned to *step_size*. (Data generated in simulations.)

very high reaching accuracy, e.g. threading a needle?

Fig. 4.9 shows a second set of error histograms for the four ITGAs with reduced camera resolution of 80x60. With all other simulation parameters held constant, the error histograms of ITGA($\tilde{f}_{arm}, \tilde{J}$) and ITGA($\tilde{f}_{arm}, J$) look much worse than their counterparts in Fig. 4.8. However, the error histogram of ITGA($f_{arm}, \tilde{J}$) does not deteriorate significantly. More interestingly, it can observed in both Fig. 4.8 and Fig. 4.9 that the reaching accuracy of ITGA($f_{arm}, \tilde{J}$) is much higher than that ITGA($\tilde{f}_{arm}, J$), which seems to indicate that the quality of $\tilde{J}$ is much better than $\tilde{f}$. Through a careful examination of the original ITGA algorithm listed in Section 4.2.3 and Eq. (4.10) and (4.11), it can be discovered that while the reaching accuracy ITGA($\tilde{f}_{arm}, J$) is

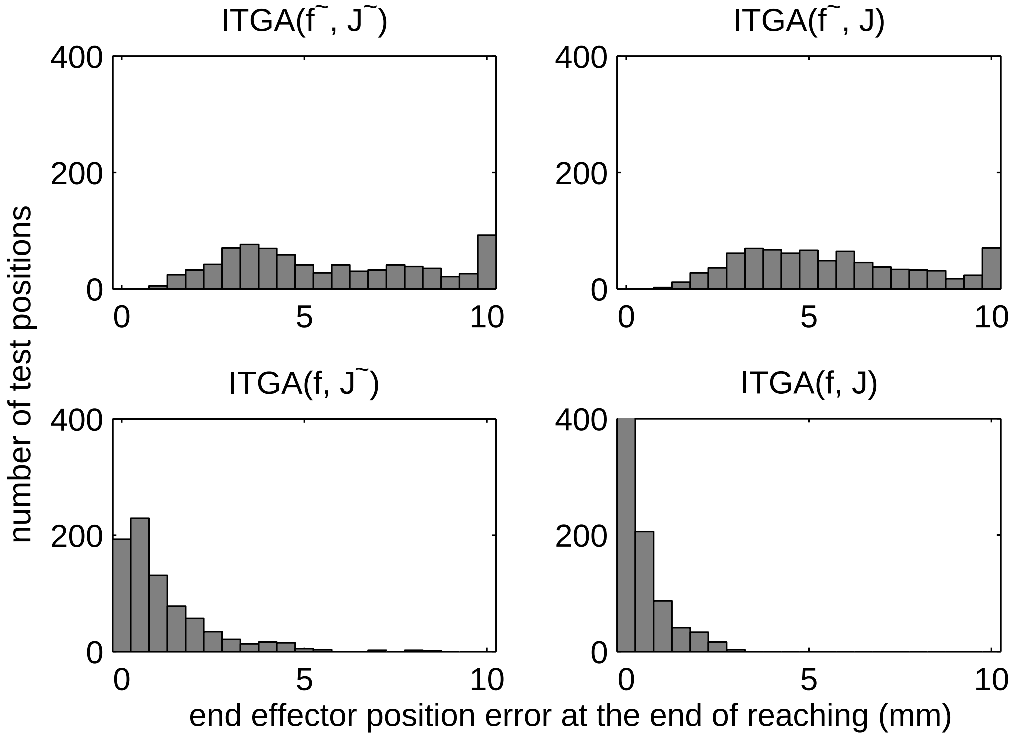


Figure 4.9: Error histograms of the four different versions of ITGA. A 80x60 camera resolution is used during motor babbling. 10mm is assigned to *step_size*. (Data generated in simulations.)

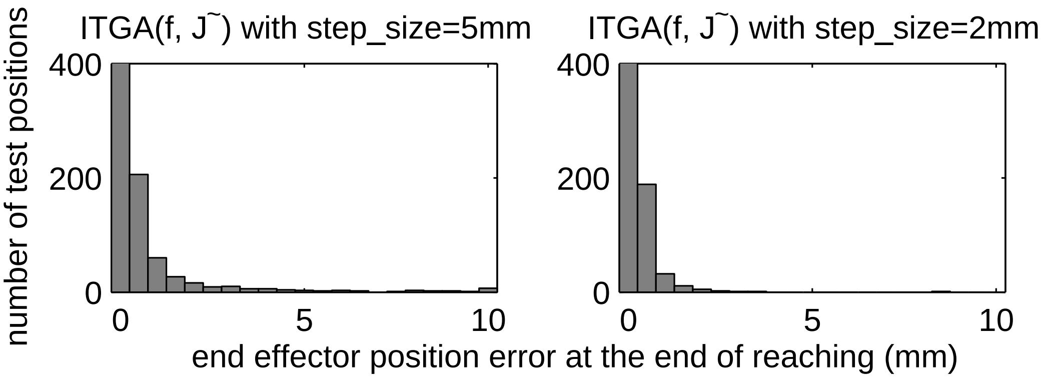


Figure 4.10: Error histograms of ITGA$(f, \tilde{J})$ based on reduced *step_size*s. A 80x60 camera resolution is used during motor babbling. (Data generated in simulations.)

almost single-handedly determined by the errors in $\tilde{f}_{arm}$, the reaching accuracy of ITGA($f_{arm}, \tilde{J}$) is determined by both the errors in $\tilde{J}$ and the variable *step_size*. Furthermore, the reaching accuracy can be improved by reducing *step_size* while keeping $\tilde{J}$ fixed. Fig. 4.10 shows two error histograms produced by ITGA($f_{arm}, \tilde{J}$) with reduced *step_size*s while inheriting the rest of the parameters from the lower left plot in Fig. 4.9. The mean reaching error achieved by ITGA($f_{arm}, \tilde{J}$) with $step\_size = 2$ is only 0.26mm, enough for threading a needle considering the average size of a needle eyelet. This result is surprising because the forward model used for Fig. 4.10 is learned with a very crude camera resolution.

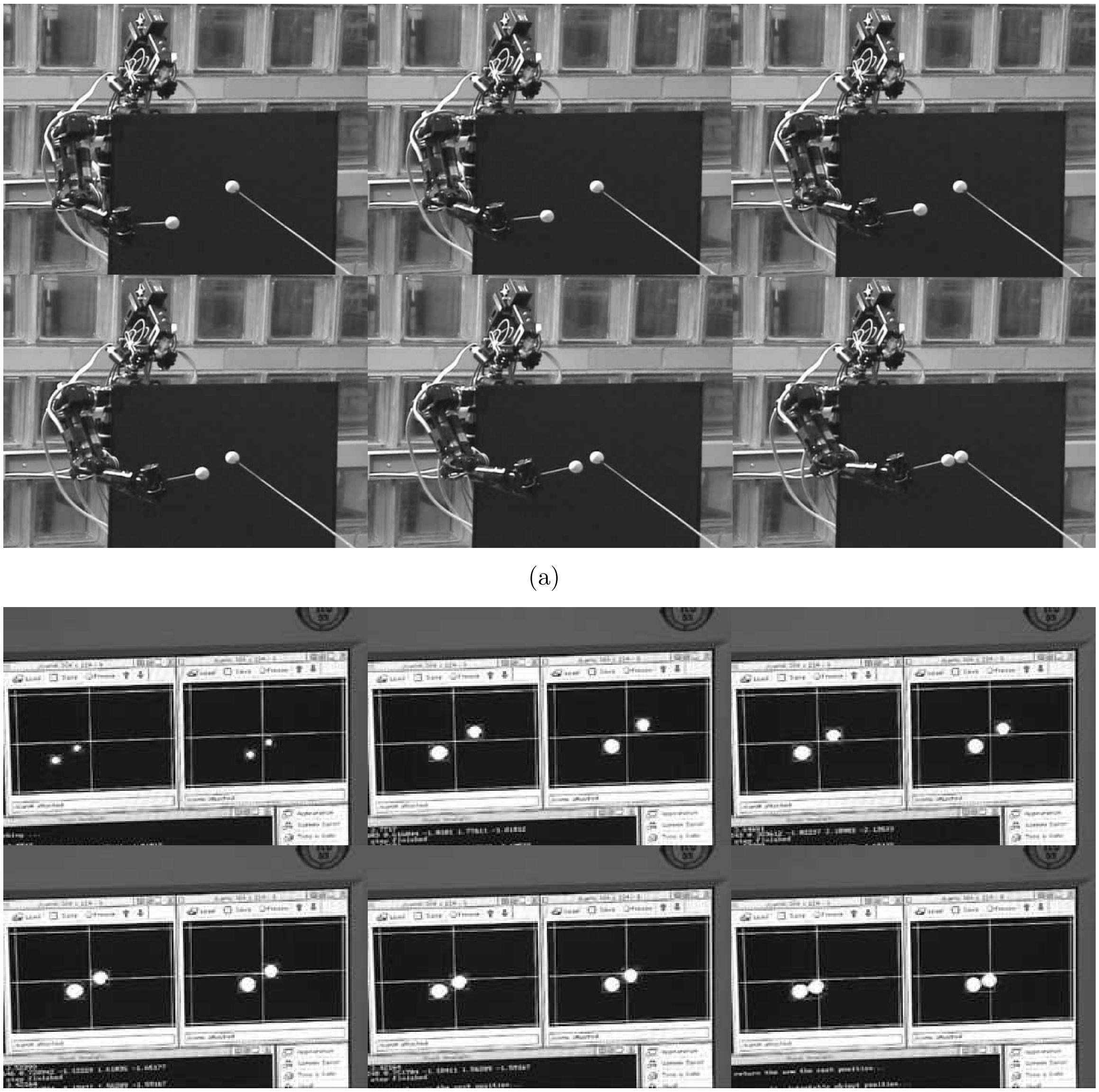
For physical experiments on a robotic platform, we have no other choice than using ITGA($\tilde{f}_{arm}, \tilde{J}$) for trajectory generation. But if we use the feedback from the stereo vision to track the position of the end effector during reaching movements, we no longer need $\tilde{f}_{arm}$ for prediction, as long as the end effector is visible to the stereo vision system. In fact, if x_{pred} is substituted by x_{perc} in Fig. 4.1, the actual reaching movements are controlled by a new ITGA. Its core equation can be described as

$$\tilde{\theta}(i) = \tilde{\theta}(i-1) + \tilde{J}^{\#}\alpha(i)(x_{target} - st(f_{arm}(\tilde{\theta}(i-1)))), \tag{4.13}$$

where the function $st()$ is the stereo perception function. The higher the resolution of visual feedback, the closer $st(x)$ is to x. With a very high camera resolution for visual feedback, the term $st(f_{arm}(\tilde{\theta}(i-1)))$ in Eq. (4.13) is virtually indistinguishable from $f_{arm}(\tilde{\theta}(i-1))$ such that the new ITGA in effect becomes ITGA($f_{arm}, \tilde{J}$). With the discussion above and the results shown in Fig. 4.10, it can be concluded that a very high reaching accuracy can be achieved with a small *step_size* and the aid of visual feedback of a high resolution. This combination can compensate for the relatively large errors in a forward model trained on samples that are gathered under a crude camera resolution.

4.3 Physical Experiments on Nico

Experiments have been carried out on Nico to test the performance of our model. A 320x240 camera resolution is used during motor babbling. 400 samples are collected to train a forward model of the arm kinematics. It takes less than 30 minutes to collect these samples. *spread* is set to 130. *margin* is determined by the heuristics described in Section 4.2.2. The reaching target is attached to the tip of a modified retractable TV antenna to facilitate its positioning. A fixed starting posture is used to generate all reaches. At the beginning of a reaching movement, *step_size* is set to 10mm and a 320x240 camera resolution is used to locate the target. After the difference between the target position and the end effector position estimated by the forward model becomes smaller than 50mm, *step_size* is reduced to 5mm and the camera resolution is switched to 640x480. Visual feedback of the end effector is exploited whenever available after the resolution switch. Fig. 4.11 shows two groups of pictures. Each group consists of six pictures organized from left to right with the pictures in the second row following the pictures in the first row. Group (A) shows a successful reach by Nico that moves its end-effector toward the target with the guidance of visual feedback. The pictures in Group (B) shows a reaching movement from the perspective of Nico's stereo vision system. Segmentation results presented in these pictures are computed using prior knowledge about color. The camera images in the first picture are captured before the resolution switch. The images in the rest of the pictures are captured with a 640x480 resolution. It can be seen from the last picture that at the end of the movement, the end effector already touches the target from the perspective of Nico's stereo vision system.



(a)

(b)

Figure 4.11: Picture sequence (a) shows Nico reaching toward a target with its end-effector. Part of Nico's body is covered with black cardboard to make it easier to identify both the target and the end-effector. Picture sequence (b) captures what the robot sees during a reaching movement. Images from both the right and left eye camera are presented in each picture. It can seen from the last picture that from the robot's perspective, the end-effector hits the target at the end of this reaching movement. Please refer to Section 4.3 for more detailed description.

4.4 Discussion

In the model proposed in this chapter for learning to reach, the only learning algorithm used is a general-purpose neural network that is widely believed to exist in the brain. While its performance can be maximized by using a carefully-designed training algorithm (OLS), simple and biological plausible Hebbian learning can used for its training as well. The iterative way the model generates reaching trajectories is consistent with the observation that infants' reaching trajectories exhibit multiple segments. The model can use visual perception of the end effector to improve reaching accuracy. It can also choose to rely on proprioception alone when the end effector is out of view or speedy trajectory generation is critical.

Previous sections have shown through both simulations and physical experiments that a very good reaching performance can be achieved with a forward model learned with a 400-sample training set. The number of hidden neurons in such a forward model is around 100. To put that into perspective, Bullock et al. used 40,000 training samples to learn the forward kinematics of a hypothetical arm with a self-organizing map containing 15625 neurons [25]. Bullock's hypothetical arm has only three DOFs and is constrained to move on a vertical plane. In contrast, our robotic arm has six DOFs and moves unconstrained in the 3D space. In Metta's work introduced in Section 2.2.2, only two of the DOFs in Babybot's arm are activated for reaching. Learning takes place through filling up a look-up table. Although the size of look-up table is not explicitly given by Metta, it grows exponentially with each additional DOF used for reaching.

In Section 4.2.3 we came to the conclusion that a very high reaching accuracy can be achieved with high resolution visual feedback and small *step_size* even when the forward model is learned with a crude visual resolution. If we assume that each reaching step is completed within a constant amount of time, this conclusion is consistent with the observation made on humans that the accuracy of a reaching movement is inversely related to the speed of the movement [47]. However, if high resolution vision can be used for feedback during reaching movements, why not use it for forward learning as well? Why bother to consider using a crude visual resolution at all? As mentioned in Section 4.2.2, because we replace the projections of the end effector on the two camera image planes with their centroids for the calculation of 3D position, the effective resolution can be crude even if a high camera resolution is used. When an infant starts to make spontaneous reaches, its vision is quite poor compared with that of an adult. The conclusion drawn in Section 4.2.3 suggests that a forward model learned with crude vision may well be used for generating accurate reaches later on when high resolution vision becomes available.

Chapter 5

Extensions of the Reaching Model

In last chapter a new model was described that allowed our humanoid robot Nico to learn to reach quickly and efficiently. The major disadvantage of this new model is that the head of the robot must maintain a fixed posture both during and after learning. If the head posture changes, learning must start all over again. In the first part of this section, an extension of the model is introduced that allows the neck joints of the robot to move freely. Instead of learning the extended kinematic chain from the eyes to the end effector directly, the sensory feedback of the vestibular system is exploited in a special way such that the dimensionality of the learning space stays unchanged. The only apparent change is that reaching trajectories that were previously straight (in Cartesian world coordinates) are now gently curved just as those produced by humans. This observation leads to the hypothesis that the gentle curvature in human reaching trajectories reflects the attempt of the brain to mitigate the degrees of freedom problem in motor learning.

While pointing involves the movement of the arm to align it with a target, it is often considered a social skill because it is used by adults to direct other people's attention. In the second part of the this chapter, another extension of the reach model is presented that allows for the production of pointing gestures without additional

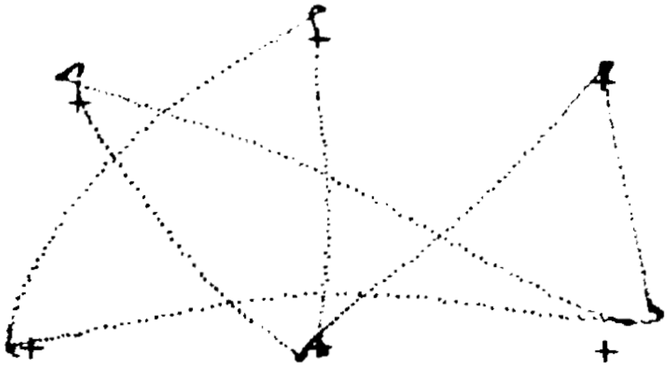


Figure 5.1: Morasso's experiments showed that when the hand moves from one planar position to another, the trajectory it follows is often mildly curved [97] The six resting positions are marked with crosshairs. The dotted curves represent some of the hand trajectories recorded during the experiments.

learning. This is achieved by manipulating the Jacobian matrix that is used for the iterative generation of motor commands. Our conjectures regarding how infants learn to distinguish between reachable and out-of-reach targets and how their pointing gestures become adult-like over time completes this chapter.

5.1 Incorporation of the Neck Joints

5.1.1 Curved Reaching Trajectories in Humans

Reaching is one of the most thoroughly investigated sensorimotor tasks. Because the human arm contains more joints than there are dimensions in the work space, there are in principle infinite postures that reach the same target in space. The experiments carried out by Morasso showed the remarkable result that human reaching movements invariably possess two characteristics: the trajectory is gently curved (see Fig. 5.1) and the velocity profile is bell-shaped [97]. Many theories have been proposed to explain the observations made by Morasso. Some are based on the assumption that reaching movements are governed by some optimization measure. For example, Uno et al. suggested that when the arm executes a reaching movement, it tries to minimizes the square of the rate of torque change over the entire movement [137]. Harris and Wolpert argued that for many types of arm movements the most important goal is to make the final position of the end effector as accurate as possible and the the stereotypical arm movements observed in humans are the result of pursuing this goal [65]. Alternative explanations of Morasso's observations include Hollerbach et al.'s hypothesis that humans plan reaching movements in the joint space and different joints start to move at different times [70]. Flash suggested that the curvature in the reaching trajectories is caused by the interactions between the viscoelastic properties of muscles and the inertial properties of the arm [49]. Wolpert et al., on the other hand, attributed the curvature to the inherent distortion in the visual perception system [150].

Guenther and Barraca's study [62] is one of the few that try to explain the curvature in reaching trajectories from the perspective of development. It is based on the work of Bullock et al. [25] described in the previous chapter. An important concept for their study is the *manipulability ellipsoid* [152], which refers to the phenomenon that if all joint increment vectors $\Delta\theta$s lying on the unit sphere are mapped into spatial displacement vectors Δxs of the arm end effector, these Δxs form an ellipsoid. The long axis of the ellipsoid is the direction along which end effector movement of unit length can be achieved with the minimum amount of joint rotation. Simulations showed that the direction mapping $\Delta x \rightarrow \Delta\theta$ learned on the data gathered during motor babbling tends to warp the moving direction of the end effector toward the long axis of the manipulability ellipsoid. Guenther and Barraca proposed that this learning bias and the tendency of the arm to move into comfortable postures contribute to the curvature in reaching trajectories observed by Morasso.

While we agree with Guenther and Barraca that the curvature in reaching trajectories can be just a side effect of motor learning, it is our belief that the cause of this side effect is more fundamental. In the next few sections, the reaching model proposed in the previous chapter will be extended to a more natural setting and we will illustrate there how a dimension reduction mechanism in the extended model results in curved reaching trajectories similar to those observed in humans.

5.1.2 Exploiting the Vestibular System

To extend the applicability of the model for learning to reach described in the previous chapter to more human-like situations, we now allow the neck joints to move freely during motor babbling. With the neck joints activated, the kinematic function mapping the joint vector into the end-effector position vector becomes: $f_{arm\_neck} : (\theta_{arm}, \theta_{neck}) \rightarrow x$. One way to cope with the expanded kinematic function is simply to increase the number of training samples. Despite the simplicity of this approach, its disadvantage is obvious. It was shown in [130] that if the two wrist joints in Nico's arm are frozen during motor babbling, about 120 training samples are needed to learn a forward model of the arm. When the two wrist joints are activated together with the shoulder and elbow joints during motor babbling, about 400 training samples are needed (see Chapter 4). In other words, the requirement for the number of training samples triples when two additional DOFs at the bottom of the arm kinematic chain are activated during motor learning. If f_{arm_neck} were to be learned directly, the increase of the training sample requirement would be even be more substantial because the neck joints are located at the top of the kinematic chain extending from the eyes to the arm end effector. Fortunately, this dramatic increase of the time to be spent on sample gathering can be avoided by taking advantage of additional sensory information.

Fig. 5.2 illustrates a simplified situation where a shift of the head posture leads to changes in both the position and the orientation of the head. The eyes are assumed to remains stationary in relation to the head. The head in its original and shifted posture is painted in black and gray respectively. Although the eye-centered coordinate system's positional change can not be perceived directly, its change in orientation can be derived from inertial information from the human vestibular system, or on Nico from a gyroscope. This information can then be used to correct this orientation shift caused by the posture change of the head, i.e. the coordinate system OXY can be corrected into $OX'Y'$ as shown in Fig. 5.2.

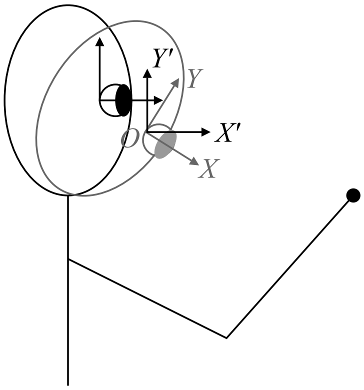


Figure 5.2: Illustration showing the effect a shifted head posture on the eye-centered coordinate system. The original and the shifted head posture is painted in black and gray respectively. Inertial information allows for the correction of the eye-centered coordination system OXY into $OX'Y'$.

The original input and output of the kinematics function f_{arm_neck} are $\theta = (\theta_{arm}, \theta_{neck})$ and

$$x = T^{eye}_{body}(\theta_{neck}) \cdot T^{body}_{ee}(\theta_{arm}) \cdot [0, 0, 0, 1]^t \tag{5.1}$$

respectively, where T^{eye}_{body} and T^{body}_{ee} are homogenous transformation matrices. T^{eye}_{body} can be expressed as

$$T_{body}^{eye} = \left[\begin{array}{cc} R_{3\times3} & T_{3\times1} \\ 0_{1\times3} & 1_{1\times1} \end{array} \right], \tag{5.2}$$

R is the rotation matrix and T is the translation vector. R can be constructed directly from sensory data from the gyroscope/vestibular system. We denote T'^{eye}_{body} as

$$T'^{eye}_{body} = \left[\begin{array}{cc} R_{3\times3} & 0_{3\times1} \\ 0_{1\times3} & 1_{1\times1} \end{array} \right]. \tag{5.3}$$

Using T'^{eye}_{body}, the perceived position x of the end-effector can be transformed into

$$x' = (T'^{eye}_{body})^{-1} \cdot x. \tag{5.4}$$

This is similar to correcting OXY into $OX'Y'$ as illustrated by Fig. 5.2.

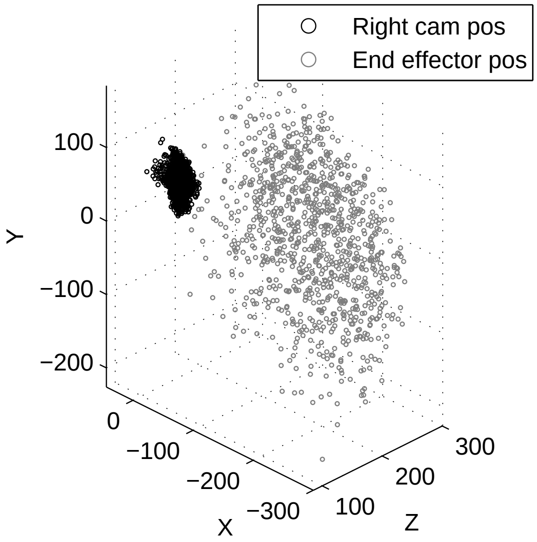


Figure 5.3: 1000 random positions for both the right eye camera (in black) and the end effector (in gray) are plotted in the body-centered coordinate system. It can be easily seen that the range of motion of the eye camera is much smaller than that of the end effector.

If we assume (falsely) that the combined effect of the neck joints is purely rotational so that it is fully eliminated in x', we can learn a forward model with the transformed training samples in the form of $((\theta_{arm})_{input}, (x')_{output})$ instead of $((\theta_{arm}, \theta_{neck})_{input}, (x)_{output})$. In reality, this assumption does not hold true, which can be easily recognized in Fig. 5.2. Learning a forward model with transformed training samples is equivalent to regarding the translational effect of the neck joints as noise. Through simulations based on the parameters of Nico, we have found that the standard deviation of the possible positions of the end effector is about three times larger than that of the eye translations caused by the neck joint rotations. Fig. 5.3 shows a scatter plot of 1000 random positions of the end-effector and 1000 random positions of the right eye camera. This difference in the ranges of motion between the end-effector and the eye camera means that the position change of the end-effector in the eye-centered coordinate system is caused to a much greater extent by the arm posture than by the head posture. The forward model learned with $((\theta_{arm})_{input}, (x')_{output})$ is not as accurate as the one learned with $((\theta_{arm}, \theta_{neck})_{input}, (x)_{output})$, but it can suffice for the purpose of reaching spatial targets. Since the body dimensions of Nico and the ranges of motion of its joints closely match those of a one-year-old child, the conclusions drawn in this paragraph also apply to a one-year-old.

The proposed learning approach which encompasses the DOFs in both the head and the arm retains the original dimensionality of problem. It has the additional advantage that the framework shown in Fig.4.1 does not need to be changed for generating reaching trajectories. Simulations show that the average positional error of blind reaching is about 20mm. Although this is significantly larger than the average positional error of the system that handles only the DOFs in the arm ($< 5mm$), it can be eliminated by exploiting visual feedback during the reaching movements. When the end effector of the arm is visible to the stereo cameras, the direction vector $\Delta\theta$ can be calculated as

$$\Delta\theta = \alpha J^{\#}(x_{target} - x_{perc}), \tag{5.5}$$

If both x_{target} and x_{perc} contain no stereo perception error, the error in $\Delta\theta$ is solely caused by the approximation error in the Jacobian J extracted from the forward model. Since $\Delta\theta$ determines the actual spatial direction the end effector moves along, the error in it causes the actual reaching trajectory to deviate from the straight line connecting the starting position of the end effector and the target.

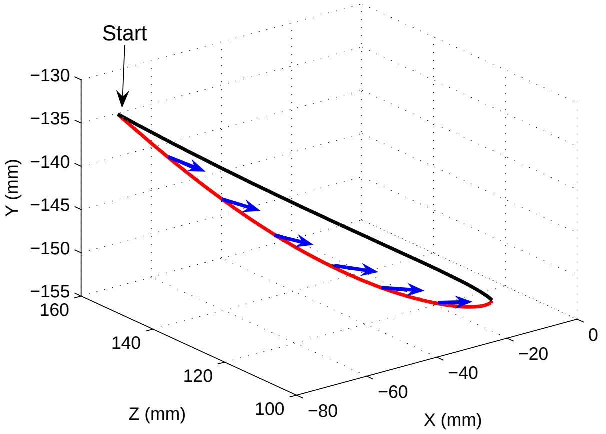


Figure 5.4: The reaching trajectory that appears to be straight is based on a forward model learned during a motor babbling phase when only the 6 DOFs in the arm are activated. The curved trajectory is based on a forward model learned on samples that use the sensory information of the gyroscope/vestibular system as a substitute for the proprioception of the neck joints. The plot is presented in the Cartesian world coordinate system. (Data generated in simulations.)

Fig. 5.4 shows two reaching trajectories produced by simulations. The trajectory that appears to be straight is generated with a forward model $\tilde{f}_{arm}$ that is learned during a motor babbling phase when all 6 DOFs in the arm are activated and the head is kept in a fixed posture. 400 samples are used to train $\tilde{f}_{arm}$. The high quality of this forward model results in a reaching trajectory that is apparently straight. The curved trajectory is generated by a forward model $\tilde{f}_{arm\_neck}$ trained on 400 samples gathered during a motor babbling phase when all DOFs in the neck and the arm are activated. These samples take the form of $((\theta_{arm})_{input}, (x')_{output})$ signifying that the sensory feedback of the gyroscope/vestibular system is employed to keep the input dimension of the forward model fixed to 6. Due to the lower quality of $\tilde{f}_{arm\_neck}$ and hence the larger error in J extracted from it, the reaching trajectory based on $\tilde{f}_{arm\_neck}$ is visibly curved. The correct reaching directions on selected points on the trajectory are marked with arrows.

5.1.3 Additional Results

To further confirm the conjecture that the curvature of the reaching trajectories based on $\tilde{f}_{arm\_neck}$ is due to the ignored translational effect of the neck joints, additional simulations have been carried out to compare the trajectories based on two different estimates of $\tilde{f}_{arm\_neck}$. For the first model, the neck joints move within their normal ranges of motion when training samples are being gathered, while for second model, each neck joint moves within a range of motion enlarged by 25%. Such an enlargement increases the average head translation during motor babbling, which in turn leads to larger errors in the second forward model. Fig. 5.5 displays two trajectories with the same starting end effector and target position. It shows the expected result that the trajectory based on the first forward model is less curved than its counterpart.

For generating the reaching trajectories shown in Fig. 5.4 and Fig. 5.5, the starting position of the end effector is chosen such that it can be perceived by the stereo vision system. The direction vector $\Delta\theta$ is always calculated according to Eq. (5.5). The variable *step_size* is assigned with a value of 1mm to achieve the best reaching accuracy. Assigning a larger value to *step_size* magnifies the error in the extracted Jacobian J and increases the curvature in the reaching trajectory as shown in Fig. 5.6. A larger *step_size* can cause the end effector to oscillate around the target before set-

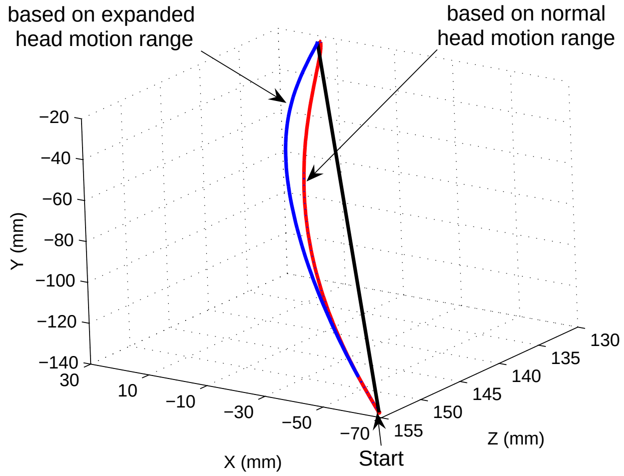


Figure 5.5: The comparison of these two trajectories show that the curvature of reaching trajecotories is indeed determined by the magnitude of the ignored head translation caused by the neck joint rotations. (Data generated in simulations.)

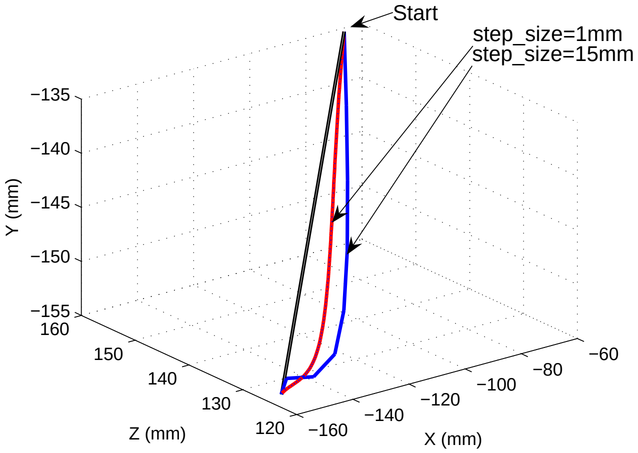


Figure 5.6: Comparison of two reaching trajectories based on different values of *step_size*. Increasing *step_size* increases the curvature of the trajectory and makes it appear less graceful. However, it also reduces the requirement for the frequency of the visual feedback. (Data generated in simulations.)

tling down. However, it also brings the benefit of reducing the requirement for the frequency of visual feedback and the computational burden associated with it. During the physical experiments on Nico, the *step_size* is initially set high and is reduced after the end effector has been moved close to the target. Fig. 5.7 shows a series

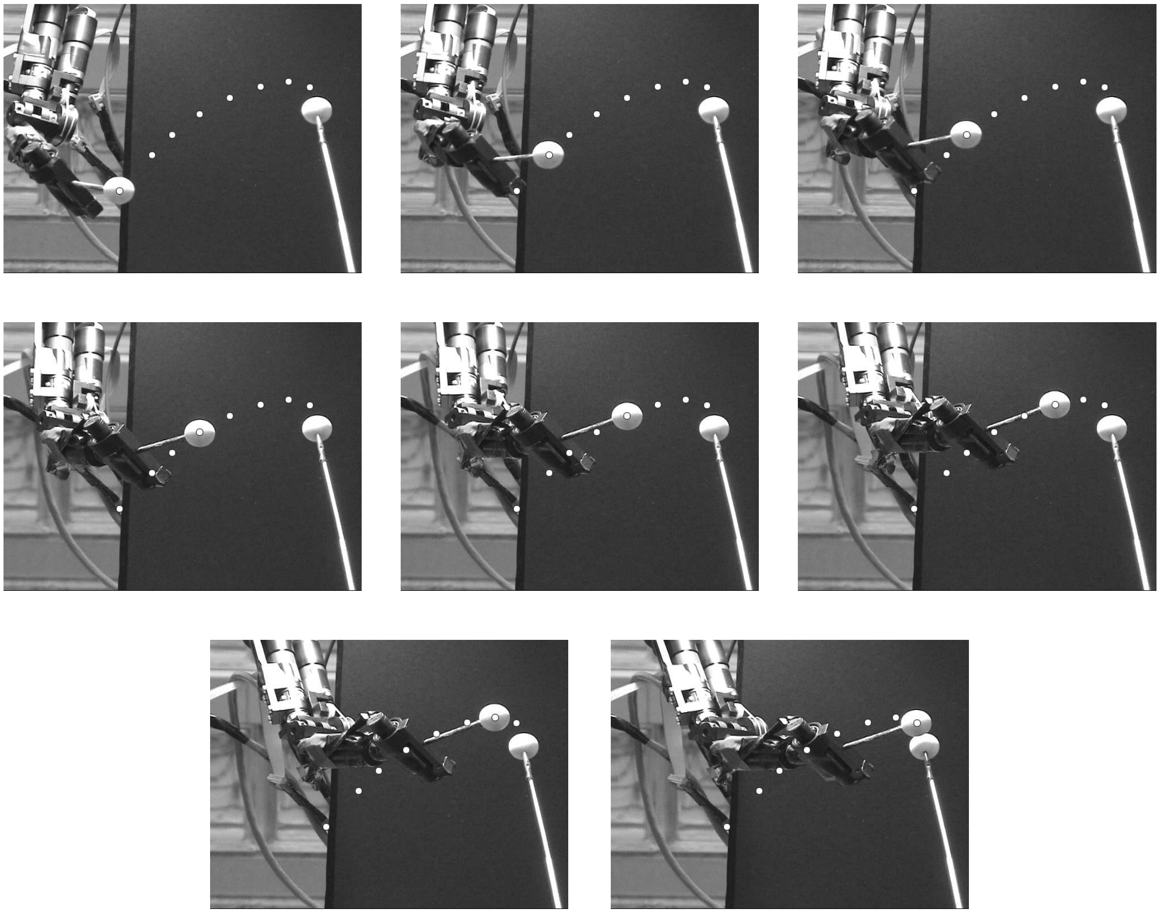


Figure 5.7: An actual reaching trajectory captured during a physical experiment on Nico. The centers of the end effector in the eight different positions on the trajectory are marked with white circles and superimposed onto the original pictures.

of pictures capturing an actual reaching trajectory during a physical experiment on Nico. A wooden ball is attached to the distal end of Nico's arm as the end effector. A wooden ball of the same size is used as the reaching target. The centers of the end effector in the eight different positions on the trajectory are marked with white circles and superimposed onto the original pictures. The trajectory looks quite natural and is only slightly more curved than the trajectories recorded by Morasso in his 1981 paper. This larger curvature can be explained by the fact that Morasso used adults for his experiments. An adult's hand has a much larger range of motion compared with his/her head. So ignoring the translational effect of the neck joints has a lesser impact on the quality of the forward model and hence results in straighter reaching trajectories. Another plausible explanation for straighter reaching trajectories observed in adults is that compared with children, adults have much more experiences with reaching so that they have acquired a better forward model.

5.1.4 Discussion

The results of both simulations and physical experiments shown in the previous sections provide support from the perspective of robotics for the feasibility of substituting neck proprioception with sensory information from the gyroscope/vestibular system during motor learning. In fact, the same substitution is used by humans for an important function - the vestibular-ocular reflex (VOR) [75]. VOR actively uses the sensory output of the vestibular nerve instead of neck proprioception to adjust eye orientations to stabilize images on the retinas. To control eye movement, both angular velocity and angular position signals are required. The vestibular nerve outputs velocity signals only. These velocity signals are integrated in the brain stem to obtain position signals. The result of this integration can be directly used to transform the position of reaching target from the eye-centered coordinate system into the body-centered coordinate system. Since no individual-specific parameters are needed for this transformation, it does not have to be learned and can be hard-coded in the genes. All simulations and physical experiments previously described are based on Nico, which is designed to match an average one-year-old. The naturally curved trajectories generated by our biologically plausible approach strongly suggest that a similar approach can be used by infants for learning to reach.

The previous sections implicitly assume that the body stays stationary relative to the outside world. If the assumption is violated, e.g. the body is passively rotated together with the head, θ_{gyro} registers a change, but the position and the orientation of the head relative to the body have not changed. Constructing the rotation matrix R in Eq. (5.3) according to the value of θ_{gyro} under this situation will lead to a reaching failure. However, this problem can be easily solved by allowing the integration of $\dot{\theta}_{gyro}$ only during the time when the neck is moving.

According to [26], there are two schemes which can be used to reach to a target. The first scheme transforms the position of the target from the eye-centered coordinate system into the body-centered coordinate system before the reaching movement is initiated. The second scheme uses the target position in the eye-centered coordinate system directly for reaching. In the previous sections, we have assumed that the first scheme is true. But the second scheme can be easily implemented with the same architecture as well. The target position x in the eye-centered coordinate system is first transformed into x' as described in Section 5.1.2. Though in a strict sense, x' is not exactly the target position in the body-center coordinate system, it can be used as a substitute for that without sacrificing much reaching accuracy. In order to be useful for the second scheme, the approach described in previous sections needs to be slightly modified such that a forward model is learned with training samples in the form of $((\theta_{arm}, \theta_{gyro})_{input}, (x)_{output})$, where θ_{gyro} is a three dimensional vector representing the sensory readings of the gyroscope/vestibular system. After training, the target position in the eye-centered coordinate system can be directly used for reaching.

It is worth noting that if samples in the form of $((\theta_{arm}, \theta_{gyro})_{input}, (x)_{output})$ are used for learning the forward model, the dimension of the input is 9, only one less than that of the input of the original f_{arm_neck}. While this seems to be only a small change in the dimensionality for our robot system, the advantage will scale with additional joints in the neck. Since the human neck has more than 4 DOFs even at the joint level, this reduction can be substantially more. On the level of muscle activation, using θ_{gyro} to represent the principle effect of the head posture delivers enormous advantage for learning to reach.

One notable paper [41] studying sensorimotor learning on a robotic platform tries to solve the degrees-of-freedom problem with a sophisticated statistical learning algorithm called Locally Weighted Projection Regression (LWPR). LWPR is used to learn the direction mapping from Δx to $\Delta\theta$ with the effort put into reducing the dimensionality locally. Our experiments suggest that the dimensionality of the problem can be reduced globally by substituting all DOFs in the neck altogether with sensory readings from the gyroscope, of which every human possess an equivalence, the vestibular system. Furthermore, the new approach described in this chapter is based on a well-studied learning algorithm (RBFN) for which there are a large number of existing implementations.

5.2 Development of Pointing

Marjanovic et al.'s work is one of the few papers in the literature that studies the learning of pointing gestures on a humanoid robot [88]. It described a two-step learning process. In the first step, a mapping from image coordinates to eye motor coordinates is learned. In the second step, a mapping from eye motor coordinates to pointing gestures is learned. After the learning, when the robot discovers an interesting object in the environment, it foveates on the object using the first mapping and then produces a corresponding pointing gesture using the second mapping. The main drawback of this approach is that it requires an artificial definition of pointing as the arm posture that makes the arm end effector cover the center the camera image.

A more intuitive definition of pointing is the arm posture that aligns the whole arm or at least the lower arm to the desired object. However, if this definition were used to learn the pointing gesture directly, the visual detection of the position and the orientation of the whole or at least part of the arm would be required. This computer vision problem is by no means easy to solve with existing techniques.

While Marjanovic et al. treated pointing as a separate skill, the Russian psychologist Vygotsky hypothesized that pointing originates from failed reaches [142]. Such reaches have the characteristics of a fully stretched arm toward objects that are too far away. According to Vygotsky, the responses of infants' caregivers to such circumstances help transform these unintentional acts into deliberate ones. (See Section 2.1.5 for more details.) The way our humanoid robot produces pointing gestures is inspired by Vygotsky's idea.

The model described in the last chapter for reaching trajectory generation is based on the following key equation:

$$\Delta\theta = J^{\#}\Delta x \tag{5.6}$$

with

$$J^{\#} = J^T(JJ^T)^{-1}. \tag{5.7}$$

When a target is too far away the elbow has to be fully extended to get as close to the target as possible. (During pointing experiments, we only activate the shoulder joints and the elbow joints since the wrist joints are not particularly useful for pointing.) After the elbow is fully extended, the 3-by-3 matrix JJ^T becomes singular because the total number of degrees of freedom of the arm degenerates into 2. Under this circumstance, $J^{\#}$ can no longer be calculated with Eq. 5.7. In order to move the arm end effector closer to the target even when J becomes singular, an extension is made regarding how $\Delta\theta$ is calculated. In the extended model, the condition number κ of J, which is the ratio of the largest to the smallest singular value of J, is continuously monitored. When κ rises above a threshold, a truncated version of J, denoted as J', is formed by cutting away the entries in J that correspond to the derivatives of x to θ_3 and θ_4 (the angles of the two elbow joints). J' is a 3-by-2 matrix. Using J', the increment of the two shoulder joint angles $\theta' = [\theta_1, \theta_2]^T$ can be calculated with

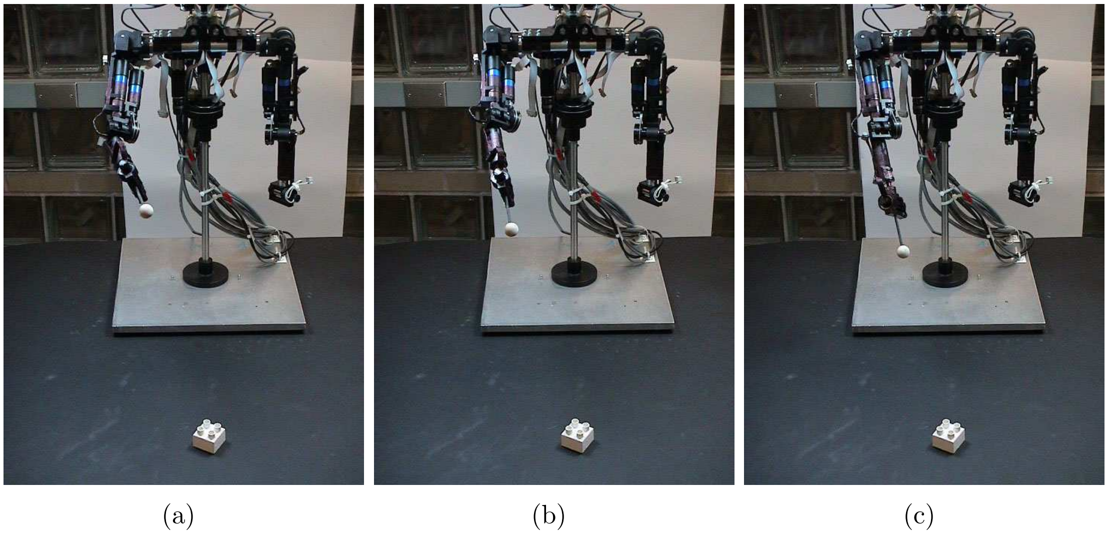
$$\Delta\theta' = J'^{\#}\Delta x, \tag{5.8}$$

where

$$J'^{\#} = (J'^T J')^{-1} J'^T. \tag{5.9}$$

$\Delta\theta'$ is then issued to the controllers of the two shoulder joints to execute another reaching step. This process is repeated until J' also becomes singular. Since the reaching model iteratively generates motor commands to minimize the distance between the arm end effector, the arm eventually moves onto the straight line that connects the shoulder base and the target. It means that the robot is in fact pointing from the shoulder base toward the target although it does not have the explicit knowledge of the vector from the origin of its camera coordinate system to the shoulder base. This knowledge would be necessary if a purely engineering approach were used for the production of pointing gestures. Fig. 5.8 shows how Nico uses the extended model to move its arm from a starting position to point toward a target it cannot reach.

As described in the previous chapter, the variable *step_size* controls the size of Δx at each reaching step and its value should be reduced when the end effector has already been moved close to the target in order to achieve high reaching accuracy.



(a) (b) (c)

Figure 5.8: A sample arm movement that leads to pointing

During the pointing experiments on Nico it has been discovered that if *step_size* maintains its initial value throughout the movement, the arm often oscillates around the straight line that connects the shoulder base and the target (see Fig. 5.9). This behavior is very similar to the imperative pointing produced by infants. In the joint attention experiments that will be described in the next chapter, experiment subjects often report that such oscillatory movements make Nico appear much more lifelike.

It is important to note that the extension just described is effectively the same as freezing the elbow joints after the elbow is fully extended and then treating the arm as a stick with two degrees of freedom at one end. This extension is also useful for certain reaching situations where the elbow is fully extended before the end effector touches the target. This can happen when the target lies at the boundary of the reachable space. Nico currently does not distinguish between objects that are reachable and objects that are too far away. But such distinction can be easily achieved by building a sample set that labels spatial positions that Nico can reach within a fixed number of steps as positive and other spatial positions as negative and then using this sample

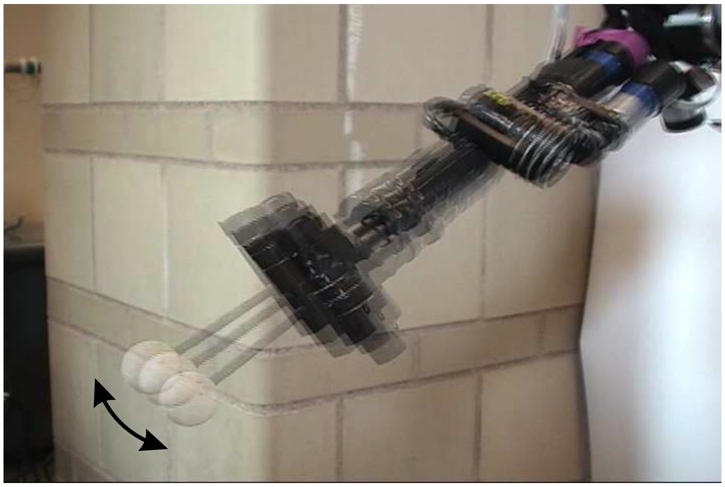


Figure 5.9: The arm often moves around the straight line that connects the shoulder base and the target if *step_size* maintains its initial value the whole time. This behavior makes Nico appear much more lifelike.

set to train a neural network.

The extended reaching model for the production of pointing gestures always makes the complete arm aligned with the target in the end. Obviously, it is not always the most energy-efficient way to point. Adults often point to a target without fully extending the elbow. How does an infant gradually learn to produce adult-like pointing gestures? Our conjecture is that the dimensions of infants' body structure play a role in this transition. Fig. 5.10 shows 25 pictures randomly captured by Nico's right camera during motor babbling. To distinguish the lower arm and the upper arm, they are marked with a piece of red and pink tape respectively. The end effector of the arm is present in all pictures. It can be seen that the lower arm is dominant in most of the pictures while the upper arm can be barely seen. When Nico points to a certain target, it keeps the end effector of the arm within its field of view. So it sees mostly the lower part of its arm after it is moved into the final configuration. Since Nico's body dimensions are about the same as those of a one-year-old, the same thing is likely to happen to infants. So when infants point to a desired object with a fully

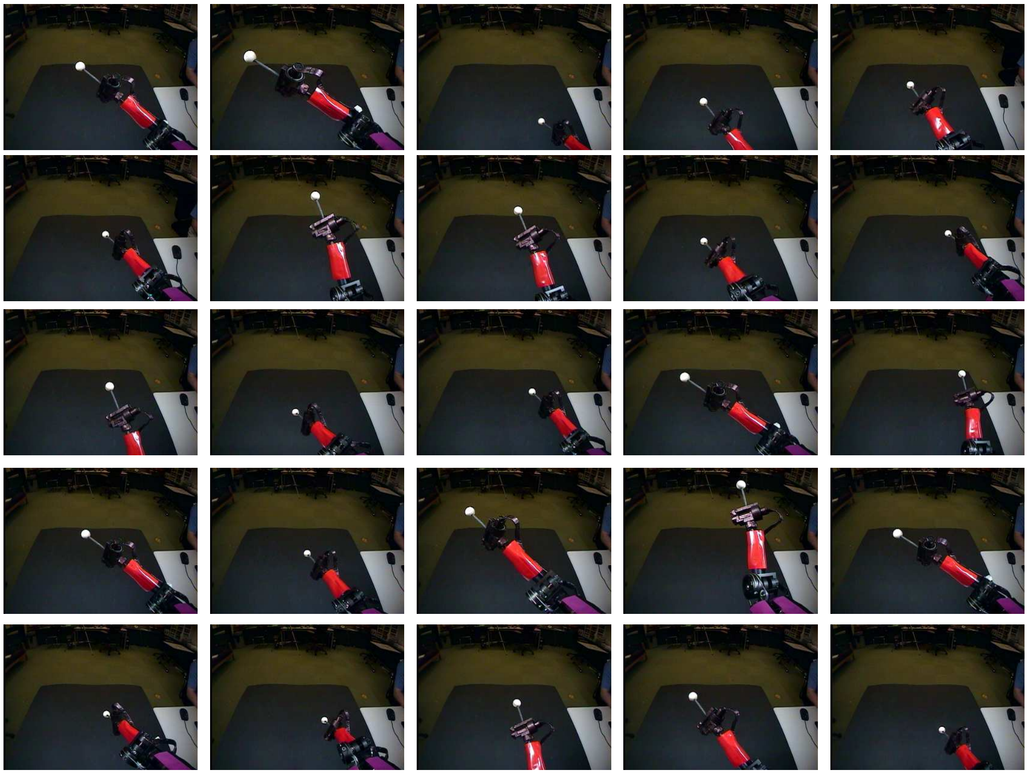


Figure 5.10: Twenty-five pictures randomly captured by Nico's right camera during motor babbling. The upper and lower arm is marked a piece of red and pink tape respectively. The lower arm is dominant in most pictures while the upper arm is barely visible.

extended arm and their caregiver retrieves the object for them, they may associate the perception of the alignment of the lower arm and the target to the reward of obtaining the object. Over time they may just stop an intended pointing movement as soon as the lower arm is already aligned with the desired object. This stop can occur when the elbow is still bent and hence results in an adult-like pointing gesture.

Chapter 6

Active Learning of Joint Attention

Joint attention is the skill of attending to the same object another person is looking at. It is only fully acquired by infants after the age of 18 months and is crucial for the development of further social and language abilities. Recently, researchers have built computational models to explain how infants learn this important skill [122][107][136]. Despite some success, most of these models are based on the assumption that the infants learn joint attention through passive observations. The models have not taken the active nature of infants and the dynamics of development into account and often require a huge number of training samples or time steps to converge. This chapter describes a system that allows our humanoid robot, Nico, to learn the joint attention skill actively and autonomously by taking advantage of existing skills and the cooperative nature of its caregiver. It results in a drastically lower requirement for training set size and a better overall performance than any other published method to date. A typical joint attention scenario between Nico and its caregiver is illustrated in Fig. 6.1.

The first section in this chapter describes the motivation of our approach. The next section presents the design of our joint attention experiments and a few technical details. The results of the experiments are given in the third section, which is followed by a discussion.

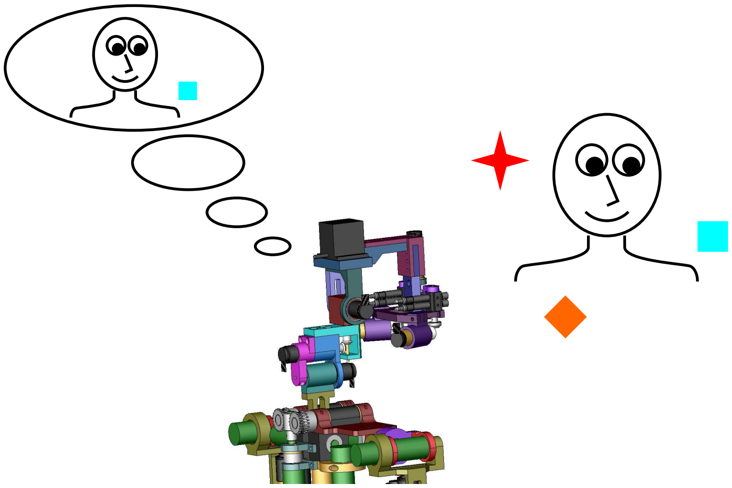


Figure 6.1: Typical joint attention scenario. The caregiver (right) is looking at an object. Nico (center) looks first at the caregiver and by judging his direction of gaze finds the object the caregiver is looking at (left).

6.1 Motivation

In Section 2.2.3, Nagai's work on simulating the development of joint attention on a humanoid robot was introduced. Two models were proposed in that work. Although both models allowed the success acquisition of joint attention, it is unlikely that they are used by infants in practice. The first model assumes that the caregiver gives the robot an explicit evaluation after it changes its gazing direction in response to her head pose change. However, there are no empirical data showing that caregivers constantly give infants right-or-wrong type evaluations during the time infants learn joint attention. In the second model, the learning of joint attention is facilitated by an innate visual attention module; explicit evaluations from the caregiver are not necessary. The problem with this model is that when there are multiple objects within the robot's field of view, the object that the innate visual attention module directs the robot's attention to coincides only by chance with the object the caregiver is looking at. This results in a large number of false samples at early stages of development when the innate visual attention module dominates the still immature joint attention module. These false samples not only slow down the learning process, but also decrease the system performance. For example, the model that employs the joint attention module needs more than 45—48×10^4 training samples to achieve to a success rate of 85% (five objects in the environment). The success rate drops quickly when the number of objects increases further.

Recently, Triesch et al. [136] proposed a computational model for the development of joint attention (or gaze following, the term preferred by the authors). The model assumes that both the infant and the caregiver are located in an idealized grid world where interesting objects can only exist at a limited number of positions (default=10). All complex behaviors of the caregiver and the infant in the real world are abstracted into a number of model parameters. Learning occurs though SARSA, a special form of reinforcement learning. The infant receives a reward when it looks at a position occupied by either the caregiver or an interesting object. The magnitude of the reward diminishes over time if the infant keeps staring at the same position and the objects in the grid world randomly move into new positions. Triesch et al. showed that in an environment as simple as has just been described, joint attention eventually emerges when the model parameters are properly set. They also postulated that the failure of children with certain developmental disorders to develop the joint attention skill can be attributed to anomalies in the parameter structure. Despite some successes, Triesch et al.'s model, similar to Nagai's model, requires a large number of time steps (on the scale of 10^4) to converge. Its performance also deteriorates quickly with the increase of the total number of objects in the grid world.

By taking a closer look at Nagai's second model and Triesch et al.'s model, it can be seen that in both models the robot/infant learns joint attention by passively watching for the change of the caregiver's head pose; the caregiver does not respond to the actions of the robot/infant in any way. Obviously, infants and caregivers in the real world do not behave like that.

Almost all major theories of development psychology acknowledge the active role played by children during their development. Infants, just like older children, do not just make passive observations; they actively explore the environment with whatever means available to them. The skills that they have already acquired often facilitate the learning of new skills. Sometimes, new skills can develop out of existing skills. Caregivers in the real world closely monitor the development of infants. They not only try to actively engage infants, but also respond promptly to infants' actions, probably more so than the other way around. The work of Collis and Schaffer shows that mothers spend much of their time following their infant's gaze direction [36]. Other experiments show that infants apparently enjoy such contingent responses and are very good at detecting them (for more details regarding contingency, please refer to Section 2.1.4).

According to Butterworth [29], infants start to produce index-finger points at the age of eleven months on average. If Vygotsky's hypothesis [142] of pointing originating from reaching is true, the onset of pointing can be dated back a few months earlier. The experiments conducted by Corkum and Moore [37] showed that infants at the age of twelve months exhibit very rudimentary joint attention; their joint attention skill can be described as reliable only after the age of fifteen months. These findings suggest that pointing precedes reliable joint attention by a few months. During these months, the following scenario can occurs repeatedly: *The infant first points to an object. This salient gesture draws the attention of the caregiver and makes him/her look at the same* *object. The infant then associates the caregiver's head pose at this particular moment to the position of the object it is pointing to.* Considering (1) the active nature of infants, (2) the willingness of caregivers to respond to infants' actions contingently and (3) the ability of infants to detect such contingent responses, the scenario described above is quite plausible and its repeated occurrences will help infants to develop joint attention. The following sections test this idea on our humanoid robot.

6.2 Method

6.2.1 Experiment Design

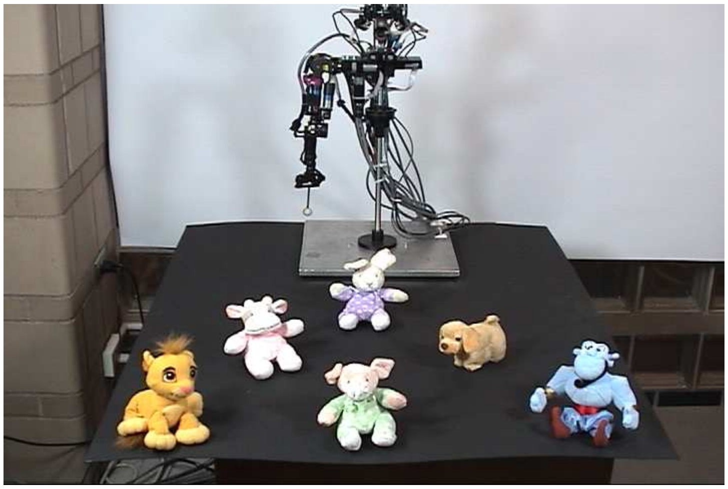


Figure 6.2: During the joint attention experiments, Nico is presented with multiple objects (stuffed animals) with salient colors lying in front of it. **Object labels:** LL: lion front left, M: pig front middle, RR: genie front right, L: cow middle left, R: dog middle right, B: rabbit in the back.

Our experiments are conducted in two phases on Nico. During the first phase, Nico learns a mapping between the head pose of its caregiver (experiment subject) and the position of the object he/she is looking at. During the second phase, Nico tests this mapping by pointing to the object it thinks its caregiver is looking at. Throughout the experiments, Nico is presented with multiple objects lying in front of it. These objects are stuffed animals with salient colors (see Fig. 6.2).

During the first phase of the experiments, Nico first engages its caregiver by looking at him/her. The robot then looks down and records the position of a random object. It then looks back at the caregiver and starts to move its arm to point toward the recorded position. The learned forward model of the arm is at this point already accurate enough to allow for the generation of pointing gestures without visual feedback. Nico maintains the pointing gesture after it is completed for a short period of time and then retracts its arm. Each of these events (start and completion of pointing, start of arm retraction) is signaled to a joint attention module. These signals are used by the joint attention module to determine when to extract an appropriate head pose of its caregiver (details regarding how this is done will be presented in Section 6.3.1). This head pose together with the position of the object Nico has just pointed to constitute a training sample. The whole process is repeated until sufficient training samples have been collected. A snapshot of this procedure is presented in Fig. 6.3.

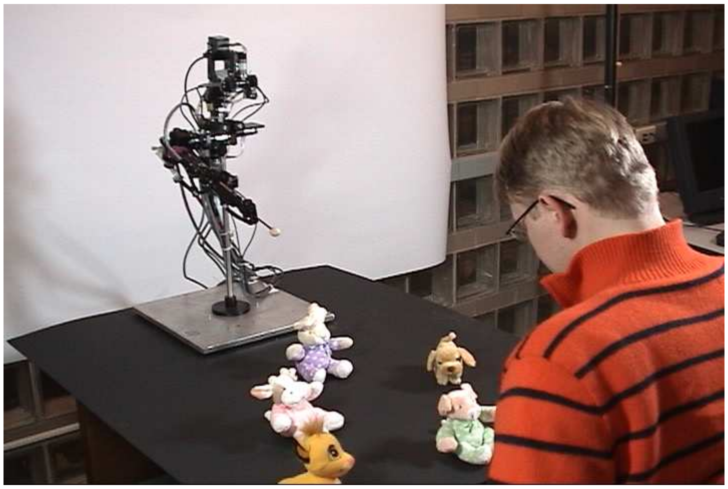


Figure 6.3: Nico points to one of the objects on the table to draw the attention of its caregiver. The caregiver reacts by looking at the object the robot points to.

In contrast to the approaches used by Nagai et al. and Triesch et al., our robot actively selects which object to attend to instead of passively watching the caregiver and trying to figure out his/her attention focus. In the process of repeatedly pointing to salient objects and watching the caregiver's responses, Nico might still collect false samples when the caregiver looks at a different object or a glitch in some component in the system causes inaccurate head pose estimation. However, it turns out that the chance of these events occurring is small.

Once sufficient training samples have been collected, a neural network is trained that maps head poses into spatial positions. The performance of the network is tested during the second phase of the experiments. In each trial, Nico continuously monitors the head pose of its caregiver. When the caregiver moves his/her head and then maintains a fixed head pose for longer than a predetermined threshold, Nico uses the trained network to project this head pose into a spatial position and points to the object that is closest to this position. The trial is considered successful when the caregiver indicates that Nico points to the object he/she has looked at.

Multiple subjects have participated in the experiments. They include members of the Social Robotics Lab, students from the Computer Science Department and students from other parts of the university. Each subject sat in front of the robot and was told to treat Nico as if it were an infant and Nico would repeatedly point to the objects on the table. The subjects were asked to keep their body still during the experiments and to look straight at Nico at the beginning to initiate the head pose estimation module in our system. Since only the head pose was estimated, we asked the subjects to move their head to indicate their attention focus whenever possible.

6.2.2 Technical Details

Fig. 6.4 illustrates the architecture of the software system for the joint attention experiments. When Nico looks down at the objects on the table, the images from its video cameras are filtered by color filters and subsequently fed into an object detection module. The object detection module uses the information in the color-filtered images to calculate the approximate positions of the objects. The forward model for reaching and pointing is already learned and the main control module uses it to generate pointing gestures toward the objects on the table. When Nico looks at its caregiver, the images from its video cameras are fed into the Small Vision System (SVS) [78], a commercial software package for calculating depth from stereo image pairs. The output of SVS is used by Watson [98] to estimate the head pose of the caregiver. Watson is a real-time head pose tracking library provided by Louis-Philippe Morency from the MIT Vision Interface Group. It is based on a framework called Adaptive View-Based Appearance Model and can track the pose of rigid objects in real-time. A joint attention module uses the output of Watson to constructing training samples for learning a mapping from head poses to object positions. During the evaluation phase, it detects when the caregiver is possibly looking one of the objects on the table and signals the main control module to generate a corresponding pointing gesture.

The modules described above are distributed among a number of QNX, Windows and Linux machines interconnected through a Gigabit network. The complete system runs at ten frames per second.

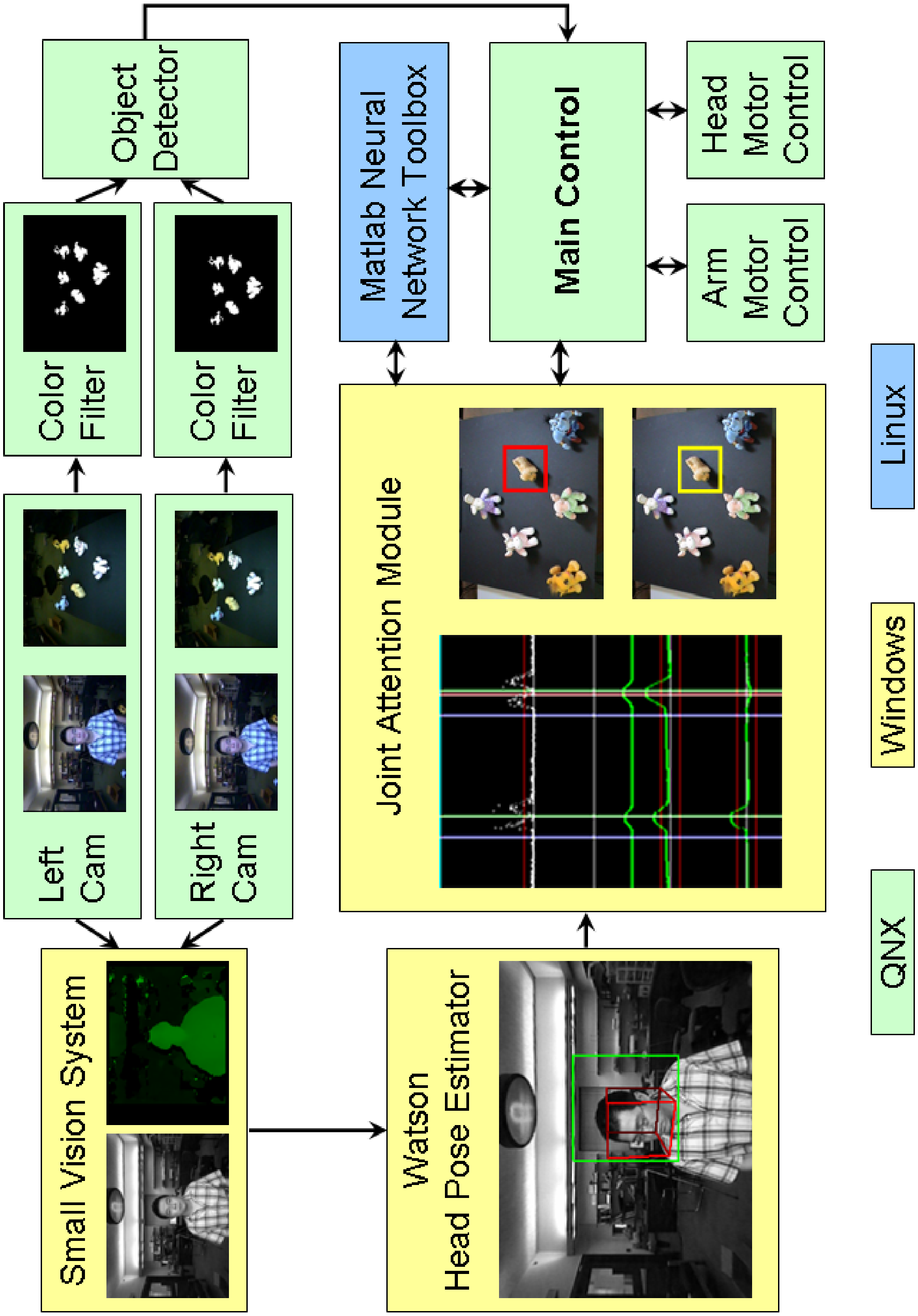


Figure 6.4: System architecture for active learning of joint attention.

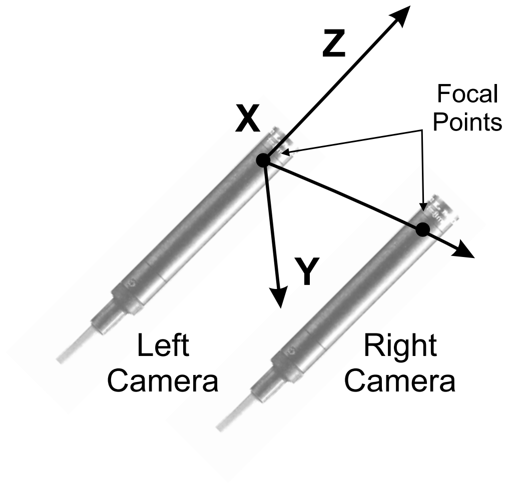


Figure 6.5: Coordinate system used to measure the positions of the objects on the table and the head poses of the experiment subjects.

The same coordinate system (illustrated in Fig. 6.5) is used to measure the position of the objects on the table and the head pose of the experiment subjects. It is based on the position of the eye cameras when Nico moves its head into the upright posture to look at its caregiver. The origin of the coordinate system is at the focal point of the left camera. The X-axis points to the focal point of the right camera. The Y-axis and Z-axis point straight down and toward the experiment subjects respectively. The approximate X and Z coordinates (in mm) of the objects on table in the coordinate system described above are as follows —

LL:	$[324, 640]^T$	M:	$[54, 640]^T$
RR:	$[-236, 640]^T$	L:	$[194, 460]^T$
R:	$[-126, 460]^T$	B:	$[54, 280]^T$

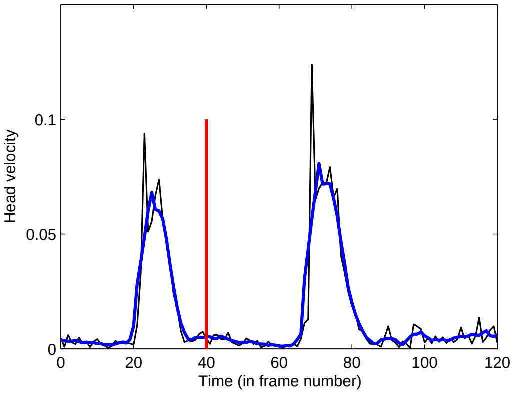
The Y components of the object positions are all of the same value because the table is parallel to the X-Z plane of the coordinate system we use. The average distance of two neighboring objects is about 250mm.

The head pose a subject maintains at the beginning of an experiment is used by Watson as the reference pose. It is described as $[0, 0, -1]^T$, which is a vector parallel to the table and pointing toward the robot. The head pose of a subject at any particular moment is described with a three dimensional vector that is relative to the reference pose. Only two of the three components in this vector are independent.

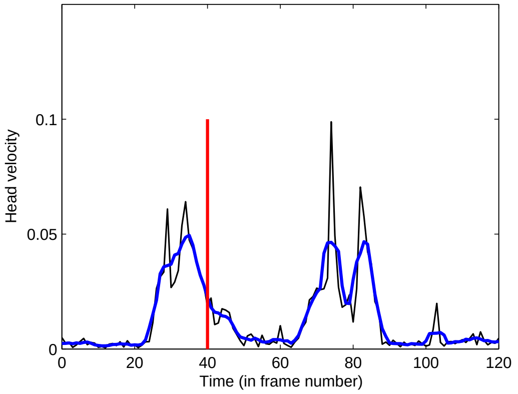
6.3 Results

6.3.1 Social Delay

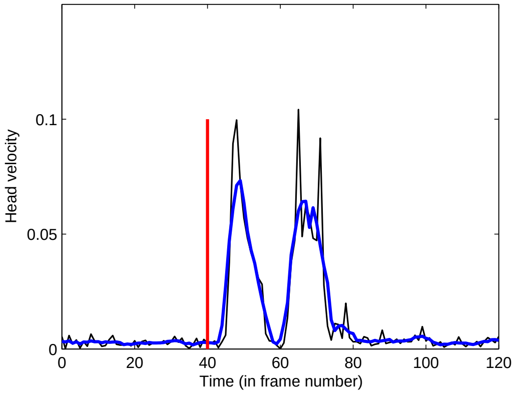
Fig. 6.6 shows examples of how three different experiment subjects respond to the robot's pointing gestures. The black curves represent the velocity of the subjects' head movement within a time window of 12 seconds (the video grabbers operate at 10 frames/second). The blue curves are simply the smoothed-out versions of the actual velocity curves using a box filter. The red line in each plot reflects the time point when the robot completes the current pointing movement. It is obvious that different subjects respond to the robot in different manners. Subject M follows the robot's arm movement closely and finds the object the robot is pointing to even before the robot fully completes the movement. The velocity curves in Fig. 6.6(a) reflect this behavior and show that the velocity of the head movement falls back to zero before the red line. Fig. 6.6(b) shows the velocity curves of Subject P who also starts to move the head soon after the robot starts the arm movement. However, P stops the head movement after the robot completes the pointing gesture. The velocity curves in Fig. 6.6(c) characterize the behavior of Subject L who apparent uses eye movements to follow the robot's arm movement and only starts to move the head after the robot completes the pointing gesture. In addition to the difference in moving the head to look at the object the robot points to, the length of the time spent looking at the object varies from person to person. While subject M often looks at the object longer than two seconds, subject L usually moves the head back quickly to look the robot in the eyes.



(a)



(b)



(c)

Figure 6.6: Examples of how three different subjects respond to the robot's pointing gestures. The black curve represents the velocity of the head movement recorded during the experiments. The blue curve shows the smoothed data. The red line marks the time when the robot completes a pointing gesture.

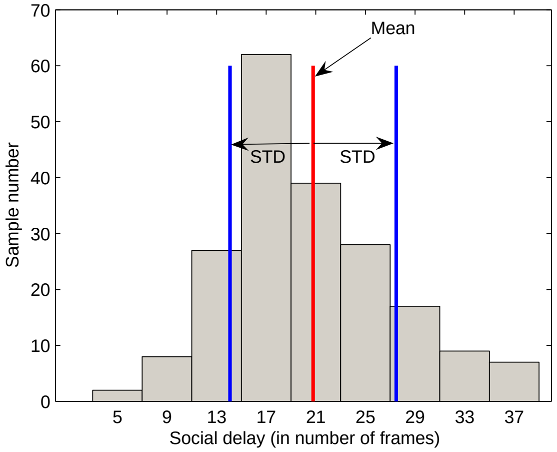


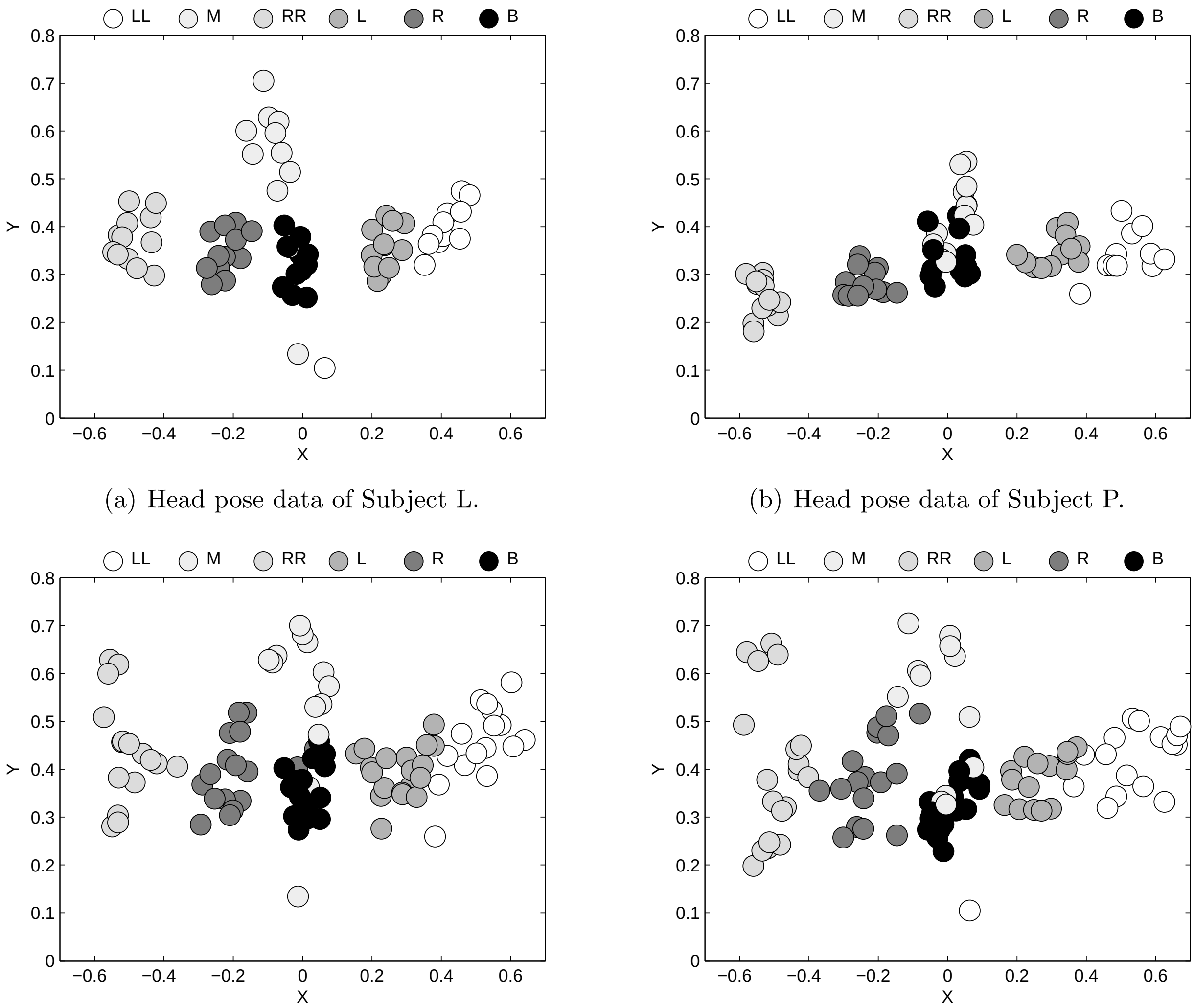
Figure 6.7: Histogram of the time delays of the experiment subjects' responses to the robot's point gestures. Despite individual differences, the distribution shows a single peak and resembles a Gaussian. The mean and the standard deviation of the distribution are $2.08s$ and $0.67s$ respectively.

During the experiments, each time the robot points to an object, the main control module needs to extract from Watson's continuous output a single head pose that best characterizes the caregiver's response. It has been hypothesized in [111] that the response delay of an individual in a social interaction can be modeled by a Gaussian distribution. A careful analysis of the data collected during our experiments has partially confirmed this hypothesis. For each of the subjects, the original velocity curves recorded during experiments are smoothed with a box filter that is twenty frames wide (2 seconds). These curves are then segmented into a number of episodes, each of them containing the subject's response to a particular pointing gesture produced by the robot. For each episode, the difference between the time the robot completes the arm movement and the time the subject exhibits a minimum amount of head movement is calculated. This difference characterizes the delay of the subject's response to the robot's social gesture. These delays are then aggregated and plotted in Fig. 6.7. It can be seen that despite the difference in the ways subjects react to the robot's pointing (as illustrated in Fig. 6.6), the distribution of the social delays shows a single peak and some resemblance to a Gaussian with a mean of $2.08s$ and a standard deviation of $0.67s$. During the automatic sample extraction mode, the main control module records the head pose data obtained from Watson after a $2s$ delay and associates this head pose with the position of the object the robot is currently pointing to.

6.3.2 System Performance

All training samples for joint attention learning were collected using the social delay approach described in the last section. Each training sample is in the form of (p_i, o_i), where p_i and o_i describe the head pose and the associated object position respectively. We use a simple Radial Basis Function Network (RBFN) to learn the association between p_i (input) and o_i (output). The two free parameters for training a RBFN - the spread of the Gaussians in the hidden layer and the error threshold as stopping criterion - are determined by a simultaneous optimization procedure.

The performance of our system is first tested by using the data collected on one subject to train a network and then evaluating the network on the same subject. The training set and the test set consist of 80 and 20 samples respectively. The performance of the trained network is very good in this case. An average recognition rate of 95% is achieved (for six objects in the environment). However, if we use a network trained on one subject directly to test on another subject's head poses, the average recognition rate is only 62%. The reason for this severe degradation is the variance of head pose data among different subjects. Fig. 6.8(a) and Fig. 6.8(b) visualize the head pose data of subject L and subject P by using only the first two components of p_i. (The third component is redundant since p_i is a normalized vector.) The shading of each marker indicates which object position it is associated with.



(a) Head pose data of Subject L.

(b) Head pose data of Subject P.

(c) A mixed data set from five different subjects (20 samples each) for training .

(d) Another mixed data set from the same subjects (20 samples each) for testing.

Figure 6.8: Head pose samples collected autonomously by the robot. Each marker represents a head pose vector projected to the X and the Y axis. The shading of each marker indicates the object position it is associated with. Experiment setting: Six objects are put at fixed positions on the table (see Fig. 6.2). Their coordinates are listed in Section 6.2.2

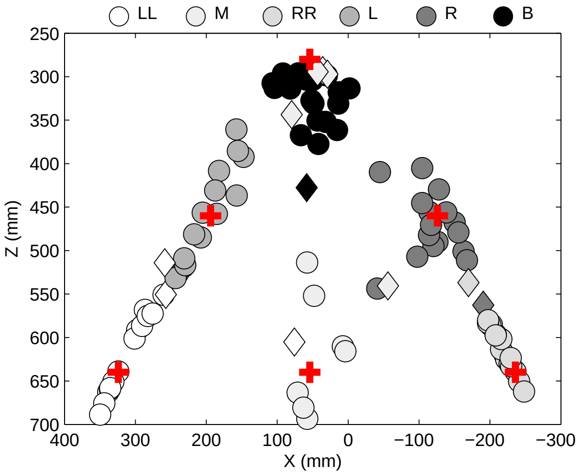


Figure 6.9: A RBFN is trained on a mixed data set portrayed in Fig. 6.8c. The performance of this RBFN is tested on another mixed data set consisting of 100 samples. This plot shows the projection of the head pose data in the test set on the X-Z plane. The original positions of the objects on the table (illustrated in Fig. 6.2) are marked with red crosshairs. The shading of each marker is determined by the closest-neighbor rule. Only 15 samples are misclassified. These samples are plotted with diamond markers.

These two plots show that although both L's and P's head pose data are well clustered, they have significant differences. These differences result in the network trained on L failing for P. However, this performance issue can be resolved by training a neural network on a mixed data set collected on different subjects. Fig. 6.8(c) shows the head pose data of a training set (100 samples) created by mixing data collected from five different subjects. When a network trained on this training set is applied to the head poses contained in another mixed data set (100 samples), a recognition rate of 85% is achieved. Fig. 6.9 shows the projection of the head poses in the test set. If a head pose is misclassified by the network, its projected position is plotted with a diamond marker instead of a round one. The axes of Fig. 6.9 are arranged in such a way that the original positions of the objects (marked with crosshairs) are topologically consistent with Fig. 6.2.

6.4 Discussion

In this chapter, we have presented a complete system for online learning of joint attention on the humanoid robot Nico. Nico applies its existing pointing skill to actively direct the attention of its caregiver. Each time it completes a pointing gesture, it captures the head pose of its caregiver after an experimentally determined time delay and associates this head pose to the position of the object it points to. A high quality training set can be built in a short amount of time by repeating this process (less than 30 minutes for 100 samples). The training set is then used to train a neural network that maps head poses into spatial positions. The performance of this network can be tested by switching the system into the evaluation mode.

A comparison of our system to Nagai's and Triesch et al.'s models for learning joint attention is presented in Table 6.1. In contrast to the two existing models, the main advantages of our system are that both learning and evaluation occur online and that far fewer training samples are required. The reason why our system works so well is that it exploits both the existing skills and the social dynamics between infant/robot and its caregiver. Nagai's and Triesch et al.'s models both assume that infants passively watch the head pose of their caregiver for clues of where he/she is looking at while the caregiver's behavior is complete independent of the infants' actions. On the contrary, our robot Nico learns joint attention by repeatedly pointing to some object that it is interested in and recording the contingent response of its caregiver. This effectively transforms the problem from extracting statistical information from data of unknown structure to learning from data that are highly structured. It is this transformation that reduces the training sample requirement of our system by several magnitudes. The drastically reduced training sample requirement also allows us to examine the differences in the response of the subjects to Nico's pointing gestures.

Currently, we rely on head pose alone to predict the attention focus of the caregiver. In the future, eye tracking support will be added to make the joint attention model more accurate. It could be achieved by first cleverly manipulating the output of Watson to determine the rough locations of the eyes and then applying template matching in the neighborhood of these locations. Another extension of the current system is to learn pointing gesture recognition. For this purpose, the learned joint attention mapping can be used to estimate the spatial position the caregiver is looking at. A new training sample can be created by combining this position vector and a feature vector that describes the pointing gesture displayed by the caregiver at the same moment.

Table 6.1: Comparison of some existing models for learning joint attention.

	Nagai's model	**Triesch et al.'s model**	**Our system**
Real world model	Yes.	No.	Yes.
Learning approach	Off-line learning.	Reinforcement learning in simulation.	Online learning.
Number of training samples/time steps required	45-48 × 10^4 samples.	About 10 × 10^4 time steps.	About 100 samples.
Recognition rate with multiple objects in view	About 85% for five objects. Drops quickly with increasing object number.	Less than 40% for six objects. Drops quickly with increasing object number.	85-95% for six objects.
Cross-subject evaluation	None. The model takes facial images as input and is therefore highly personalized.	None. Caregivers are characterized only by their probability of looking at a salient object at any given moment.	Works well when the training set contains samples from multiple subjects.

Chapter 7

Conclusion

A completely new humanoid robot named Nico that matches the body dimensions of an average one-year old infant has been designed and assembled by the author of this thesis over the course of the past few years. It has a total of 21 degrees of freedom, each individually driven by a miniature DC motor. Four micro CCD-cameras, one gyroscope and the built-in optical encoders of the motors constitute the sensory system of the robot. The software system of the robot has been co-developed by the author and his colleagues. Not only has Nico served well as the platform for the experiments in this thesis, it has also been used for the other studies such as self-other discrimination, human perception of humanoid robot behaviors, pronoun learning, to name just a few.

In this thesis, the development of reaching, pointing and joint attention has been implemented on Nico to demonstrate the power of skill progression, an important feature of development learning that has not been paid proper attention to in the field of humanoid robotics. These three skills were chosen partially because they have all been extensively studied by developmental psychologists. And outside the field of development psychology, physiologists and neuroscientists have been trying for years to uncover the neural basis of reaching and pointing. The large number of past observations and hypotheses on the development of these skills in infants have provided both inspirations and a ground for comparison. Although there are no lack of learning methods for humanoid robots to acquire these skills, today on most humanoid robots, these skills are still implemented through careful engineering. Part of the reason for that is that previously published learning methods are often either very slow or difficult to implement, and can contain very specific assumptions. These are the problems we have tried to overcome in this thesis.

The model for reaching introduced in Chapter 3 requires only a calibrated stereo camera system and the ability to localize the arm end effector. It uses a Radial Basis Function Network (RBFN), a general-purpose neural network, to approximate the forward kinematics of Nico's 6-DOF arm. When the two training parameters of the RBFN are properly tuned, only 400 training samples are needed that can be collected within 30 minutes. If visual feedback of the end effector is available during reaching, the accuracy of the reaching movements generated by the model is only limited by the highest visual resolution available. Chapter 4 describes two extension of the reaching model. The first extension allows the neck joints of the robot to move freely during motor babbling. The second extension enables the robot to produce pointing gestures toward distant targets. Both extensions require no additional training samples. The learning of joint attention takes advantage of the pointing skill and the cooperative nature of the caregiver. The underlying software system contains a commercially available stereo vision system and a head pose estimation system (that can actually estimate the pose of any rigid object) free for educational and research purposes. Using this system one hundred training samples can be collected within 30 minutes. The resulting joint attention model, again, represented with a RBFN, performs at a success rate of 85%.

The previous paragraph has shown that the development of reaching, pointing and joint attention on Nico requires little more than a few basic visual perception abilities and a general learning algorithm (RBFN). However, because it has taken full advantage of the intimate connections among these skills, the complete development takes only about one hour in total (30 minutes for reaching and pointing + 30 minutes for joint attention). Only 500 training samples need to be collected during the process (400 for reaching and pointing + 100 for joint attention). This number is a few magnitudes smaller than the one required by previous works just for the learning of joint attention alone. Although this dramatic improvement of efficiency is unlikely to occur for any arbitrary skill learning task, it appears that some further exploration of the concept of skill progression is worthwhile.

Bibliography

[1] K. E. Adolph, B. Vereijken, and M. A. Denny. Learning to crawl. *Child Development*, 69(5):1299–1312, 1998.

[2] R. O. Ambrose, H. Aldridge, R. S. Askew, R. Burridge, W. Bluethman, M. A. Diftler, C. Lovchik, D. Magruder, and F. Rehnmark. Robonaut: Nasa's space humanoid. *IEEE Intelligent Systems*, 15:57–63, 2000.

[3] C. Amiel-Tison and A. Grenier. *Neurological assessment during the first year of life*. Oxford University Press, New York, 1986.

[4] D. I. Anderson, J. J. Campos, and M. A. Barbu-Roth. A development perspective on visual proprioception. In G. Bremner and A. Slater, editors, *Theories of infant development*, pages 30–69. Blackwell Publishing, Oxford, 2004.

[5] M. Asada, K. F. MacDorman, H. Ishiguro, and Y. Kuniyoshi. Cognitive developmental robotics as a new paradigm for the design of humanoid robots. *Robotics and Autonomous Systems*, 37:185–193, 2001.

[6] D. A. Baldwin. Understanding the link between joint attention and language. In C. Moore and P. J. Dunham, editors, *Joint Attention: Its origins and role in development*, pages 131–158. Lawrence Erlbaum Associates, Hillsdale, N.J., 1995.

[7] A. Bandura. *Social foundations of thought and action*. Prentice-Hall, Englewood Cliffs, N.J., 1986.

[8] S. Baron-Cohen. *Mindblindness: Ann essay on autism and theory of mind*. MIT Press, Cambridge, M.A., 1995.

[9] J. J. Barton, D. Z. Press, J. P. Keenan, and M. O'Connor. Lesions of the fusiform face area impair perception of facial configuration in prosopagnosia. *Neurology*, 58(1):71–78, 2002.

[10] E. Bates, L. Camaioni, and V. Volterra. The acquisition of performatives prior to speech. *Merrill-Palmer Quarterly*, 21:205–226, 1975.

[11] N. A. Bernstein. *The co-ordination and regulation of movements*. Pergamon Press, 1967.

[12] B. I. Bertenthal and D. L. Bai. Infants' sensitivity to optical flow for controlling posture. *Developmental Psychology*, 25:936–945, 1989.

[13] N. E. Berthier, R. K. Clifton, D. D. McCall, and D. J. Robin. Proximodistal structure of early reaching in human infants. *Experimental Brain Research*, 127:259–269, 1999.

[14] N. E. Berthier and R. Keen. Development of reaching in infancy. *Experimental Brain Research*, 169:507–518, 2006.

[15] N. E. Berthier and M. McCarty. Speed of infant reaching during the first year: Confirmation of a prediction. *Infant Behavior and Development*, 19:531, 1996.

[16] C. M. Bishop. *Neural networks for pattern recognition*, pages 148–150. Oxford University Press, Oxford, 1995.

[17] P. Bloom. *How Children Learn the Meanings of Words.* MIT Press, Cambridge, M.A., 2000.

[18] J.-Y. Bouguet. Camera calibration toolbox for matlab. `http://www.vision.caltech.edu/bouguetj/calib_doc/`.

[19] C. Breazeal. *Designing sociable robots.* MIT Press, Cambridge, M.A., 2002.

[20] C. Breazeal, A. Brooks, J. Gray, G. Hoffman, J. Lieberman, H. Lee, A. Lockerd, and D. Mulanda. Tutelage and collaboration for humanoid robots. *International Journal of Humanoid Robotics*, 1(2):315–348, 2004.

[21] C. L. Breazeal. *Sociable machines: Expressive social exchange between humans and robots.* PhD thesis, Massachusetts Institute of Technology, 2000.

[22] R. A. Brooks. Elephants don't play chess. *Robotics and Autonomous Systems*, 6:3–15, 1990.

[23] R. A. Brooks, C. Breazeal (Ferrell), R. Irie, C. C. Kemp, M. Marjanovic, B. Scassellati, and M. M. Williamso. Alternative essences of intelligence. In *Proceedings of the American Association of Artificial Intelligence (AAAI-98)*, 1998.

[24] R. A. Brooks and L. A. Stein. Building brains for bodies. *Autonomous Robots*, 1:7–25, 1994.

[25] D. Bullock, S. Grossberg, and F. H. Guenther. A self-organizing neural model of motor equivalent reaching and tool use by a multijoint arm. *Journal of Cognitive Neuroscience*, 5(4):408–435, 1993.

[26] C. A. Buneo, M. R. Jarvis, A. P. Batista, and R. A. Andersen. Direct visuomotor transformations for reaching. *Nature*, 416:632–636, 2002.

[27] N. J. Butko, I. R. Fasel, and J. R. Movellan. Learning about humans during the first 6 minutes of life. In *Proceedings of the 5th IEEE International Conference on Development and Learning*, 2006.

[28] G. Butterworth and N. Jarrett. What minds have in common is space: Spatial mechanisms serving joint visual attention in infancy. *British Journal of Developmental Psychology*, 9:55–72, 1991.

[29] G. E. Butterworth. Joint visual attention in infancy. In G. Bremner and A. Fogel, editors, *Blackwell handbook of infant development*, pages 213–240. Blackwell Publishing, Oxford, 2001.

[30] G. E. Butterworth and S. Itakura. How the head, eyes and hands serve definite reference. *British Journal of Developmental Psychology*, 18:25–50, 2000.

[31] G. E. Butterworth and P. Morissette. Onset of pointing and the acquisition of language in infancy. *Journal of Reproductive and Infant Psychology*, 14:219–231, 1996.

[32] S. Calinon and A. Billard. Learning of gestures by imitation in a humanoid robot. In K. Dautenhahn and C. L. Nehaniv, editors, *Imitation and social learning in robots, humans and animals: behavioural, social and communicative dimensions*. Cambridge University Press, Cambridge, 2006.

[33] L. Camaioni. The development of intentional communication: A re-analysis. In J. Nadel and L. Camaioni, editors, *New perspectives in early communicative development*, pages 82–96. Routledge, London, 1993.

[34] S. Chen, C. F. N. Cowan, and P. M. Grant. Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Transactions on Neural Networks*, 2(2):302–309, 1991.

[35] R. K. Clifton, D. W. Muir, D. H. Ashmead, and M. G. Clarkson. Is visually guided reaching in early infancy a myth? *Child Development*, 64:1099–1110, 1993.

[36] G. M. Collis and H. R. Schaffer. Synchronization of visual attention in mother-infant pairs. *Journal of Child Psychology and Psychiatry*, 16:315–320, 1975.

[37] V. Corkum and C. Moore. Development of joint visual attention in infants. In C. Moore and P. J. Dunham, editors, *Joint Attention: Its origins and role in development*, pages 61–83. Lawrence Erlbaum Associates, Hillsdale, N.J., 1995.

[38] J. Decety, T. Chaminade, J. Grezes, and A. N. Meltzoff. A pet exploration of the neural mechanisms involved in reciprocal imitation. *Neuroimage*, 15:265–272, 2002.

[39] D. DeMers and K. Kreutz-Delgado. Inverse kinematics of dextrous manipulators. In O. Omidvar and P. van der Smagt, editors, *Neural systems for robotics*, pages 75–116. Academic Press, New York, 1997.

[40] S. Deneve, P. E. Latham, and A. Pouget. Efficient computation and cue integration with noisy population codes. *Nature Neuroscience*, 4(8):826–831, 2001.

[41] A. D'Souza, S. Vijayakumar, and S. Schaal. Learning inverse kinematics. In *Proceedings of the 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2001.

[42] A. Edsinger-Gonzales and J. Weber. Domo: A force sensing humanoid robot for manipulation research. In *Proceedings of the 4th IEEE-RAS/RSJ International Conference on Humanoid Robots*, 2004.

[43] N. J. Emery, E. N. Lorincz, D. I. Perret, M. W. Oram, and C. I Baker. Gaze following and joint attetnion in rhesus monkeys (*Macaca mulatta*). *Journal of Comparative Psychology*, 111:1–8, 1997.

[44] M. J. Farah, C. Rabinowitz, G. E. Quinn, and G. T. Liu. Early commitment of neural substrates for face recognition. *Cognitive Neuropsychology*, 17(1-3):117–123, 2000.

[45] T. Farroni, E. M. Mansfield, C. Lai, and M. H. Johnson. Infants perceiving and acting on the eyes: Tests of an evolutionary hypothesis. *Journal of Experimental Child Psychology*, 85:199–212, 2003.

[46] L. Fetters and J. Todd. Quantitative assessment of infant reaching movements. *Journal of Motor Behavior*, 19:147–166, 1987.

[47] P. M. Fitts. The information capacity of the human motor system in controlling the amplitude of movements. *Journal of Experimental Psychology*, 47:381–391, 1954.

[48] P. M. Fitzpatrick. *From first contact to close encounters: A developmentally deep perceptual system for a humanoid robot.* PhD thesis, Massachusetts Institute of Technology, 2003.

[49] T. Flash. Generation of reaching movements: Plausibility and implications of the equilibrium trajectory hypothesis. *Brain, Behavior and Evolution*, 33(2-3):63–68, 1989.

[50] A. Fod, M. J. Mataric, and O. C. Jenkins. Automated derivation of primitives for movement classification. *Autonomous Robots*, 12(1):39–54, 2002.

[51] J. A. Fodor. *The modularity of mind.* MIT Press, Cambridge, M.A., 1983.

[52] R. Fox and C. McDaniel. The perception of biological motion by human infants. *Science*, 218:486–487, 1982.

[53] F. Franco and G. E. Butterworth. Pointing ad social awareness: Declaring and requesting in the second year of life. *Journal of Child Language*, 23(2):307–336, 1996.

[54] G. Gallup, J. Anderson, and D. Shillito. The mirror test. In M. Bekoff, C. Allen, and G. Burghardt, editors, *The cognitive animal: empirical and theoretical perspectives on animal cognition*, pages 325–333. MIT Press, Cambridge, M.A., 2002.

[55] I. Gauthier, M. J. Tarr, A. W. Anderson, P. Skudlarski, and J. C. Gore. Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. *Nature Neuroscience*, 2(6):568–573, 1999.

[56] E. J. Gibson. *Principles of perceptual learning and development.* Appleton-Century-Crofts, New York, 1969.

[57] E. J. Gibson. The concept of affordances in development: The renascence of functionalism. In W. A. Collins, editor, *The concept of development*, pages 55–81. Erlbaum, Hillsdale, N.J., 1982.

[58] E. J. Gibson and R. D. Walk. The 'visual cliff'. *Scientific American*, 202:64–71, 1960.

[59] F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural networks architectures. *Neural Computation*, 7:219–269, 1995.

[60] S. F. Giszter, F. A. Mussa-Ivaldi, and Emilio Bizzi. Convergent force fields organized in the frog's spinal cord. *The Journal of Neuroscience*, 13(2):467–491, 1993.

[61] C. C. Goren, M. Sarty, and P. Y. K. Wu. Visual following and pattern discrimination of face-like stimuli by newborn infants. *Pediatrics*, 56:544–549, 1975.

[62] F. H. Guenther and D. M. Barreca. Neural models for flexible control of redundant systems. In P. G. Morasso and V. Sanguineti, editors, *Self-organization, computational maps, and motor control*, pages 383–421. Elsevier, 1997.

[63] S. M.J. Hains and D. W. Muir. Effects of stimulus contingency in infant-adult interactions. *Infant Behavior and Development*, 19:49–61, 1996.

[64] J. D. Han, S. W. Zeng, K. Y. Tham, M. Badgero, and J. Weng. Dav: A humanoid robot platform for autonomous mental development. In *Proceedings of the 2nd International Conference on Development and Learning*, 2002.

[65] C. M. Harris and D. M. Wolpert. Signal-dependent noise determines motor planning. *Nature*, 394:780–784, 1998.

[66] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning.* Springer Verlag, New York, 2001.

[67] J. B. Haviland. How to point in zinacantan. In S. Kita, editor, *Pointing: Where language, culture and cognition meet*, pages 139–170. Lawrence Erlbaum Associates, Mahwah, N.J., 2003.

[68] S. Haykin. *Neural networks: A comprehensive foundation.* Prentice Hall, Upper Saddle River, N.J., second edition, 1999.

[69] J. Heikkilä and O. Silvén. Calibration procedure for short focal length off-the-shelf ccd cameras. In *Proceedings of the 1996 International Conference on Pattern Recognition*, 1996.

[70] J. M. Hollerbach, S. P. Moore, and C. G. Atkeson. Staggered joint interpolation as an underlying coordination strategy during human arm movement. In *Proceedings of International Union of Physiological Sciences, Vol. XVI*, 1986.

[71] A. Ijspeert, J. Nakanishi, and S. Schaal. Learning attractor landscapes for learning motor primitives. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 1547–1554. MIT Press, Cambridge, M.A., 2003.

[72] M. H. Johnson and A. Karmiloff-Smith. Neuroscience perspectives on infant development. In G. Bremner and A. Slater, editors, *Theories of Infant Development*, pages 121–141. Blackwell Publishing, Oxford, 2004.

[73] S. Johnson, V. Slaughter, and S. Carey. Whose gaze will infants follow? the elicitation of gaze following in 12-month-olds. *Developmental Science*, 1(2):233–238, 1998.

[74] M. I. Jordan and D. Rumelhart. Supervised learning with a distal teacher. *Cognitive Science*, 16:307–354, 1992.

[75] Eric R. Kandel, James H. Schwartz, and Thomas M. Jessell. *Principles of Neural Science.* McGraw-Hill, New York, fourth edition, 2000.

[76] A. Karmiloff-Smith. The connectionist infant: Would piaget turn in his grave? *SRCD Newsletter, Fall*, pages 1–3, 1996.

[77] J. Konczak and J. Dichgans. The development toward stereotypic arm kinematics during reaching in the first 3 years of life. *Experimental Brain Research*, 117:346–354, 1997.

[78] K. G. Konolige. Small vision system. `http://www.ai.sri.com/software/SVS`.

[79] H. Kozima, C. Nakagawa, D. Kosugi, N. Kawai, and Y. Yano. A humanoid robot in company with children. In *Proceedings of the 4th IEEE-RAS/RSJ International Conference on Humanoid Robots*, 2004.

[80] H. Kozima, C. Nakagawa, and Y. Yasuda. Designing and observing human-robot interactions for the study of social development and its disorders. In *Proceedings of 2005 IEEE International Symposium on Computational Intelligence in Robotics and Automation*, 2005.

[81] H. Kozima, C. Nakagawa, and Y. Yasuda. Interactive robots for communication-care: A case-study in autism therapy. In *Proceedings of 2005 IEEE International Workshop on Robots and Human Interactive Communication*, 2005.

[82] A. M. Leslie. The theory of mind impairment in autism: evidence for a modular mechanism of development? In A. Whiten, editor, *Natural theories of mind: Evolution, development and simulation of everyday mindreading*, pages 63–78. Basil Blackwell, Oxford, 1991.

[83] A. M. Leslie. Tomm, toby, and agency: Core architecture and domain specificity. In L. A. Hirschfeld and S. A. Gelman, editors, *Mapping the mind: Domain*

specificity in cognition and culture, pages 119–148. Cambridge University Press, Cambridge, 1994.

[84] A. Liegeois. Automatic supervisory control of the configuration and behavior of multibody mechanisms. *IEEE Transactions on Systems, Man and Cybernetics*, 7(12):868–871, 1977.

[85] A. Lock, A. Young, V. Service, and P. Chandler. Some observations on the origins of the pointing gesture. In V. Volterra and C. J. Erting, editors, *From gesture to language in hearing and deaf children*, pages 42–55. Springer, Berlin, 1990.

[86] M. Lungarella, G. Metta, R. Pfeifer, and G. Sandini. Developmental robotics: a survey. *Connection Science*, 15(4):151–190, 2003.

[87] A. A. Maciejewski and C. A. Klein. Obstacle avoidance for kinematically redundant manipulators in dynamically varying environments. *The international journal of robotics research*, 4(3):109–117, 1985.

[88] M. Marjanovic, B. Scassellati, and M. Williamson. Self-taught visually guided pointing for a humanoid robot. In *Proceedings of the 4th International Conference on Simulation of Adaptive Behavior*, 1996.

[89] P. V. McDonald, R. E. A. van Emmerik, and K. M. Newell. The effects of practice on limb kinematics in a throwing task. *Journal of Motor Behavior*, 21:245–264, 1989.

[90] A. N. Meltzoff and M. K. Moore. Imitation of facial and manual gestures by human neonates. *Science*, 198:75–78, 1977.

[91] D. Messer. Processes of development in early communication. In G. Bremner and A. Slater, editors, *Theories of infant development*, pages 284–316. Blackwell Publishing, Oxford, 2004.

[92] G. Metta. *Babyrobot: A study on sensori-motor development.* PhD thesis, University of Genova, 1999.

[93] R. C. Miall. Motor control, biological and theoretical. In M. A. Arbib, editor, *The handbook of brain theory and neural networks*, pages 686–689. MIT Press, Cambridge, M.A., 2002.

[94] R. C. Miall and D. M. Wolpert. Forward models for physiological motor control. *Neural Networks*, 9:1265–1279, 1996.

[95] W. S. Millar and J. S. Watson. The effect of delayed feedback on infant learning reexamined. *Child Development*, 50:747–751, 1979.

[96] P. H. Miller. *Theories of developmental psychology.* Worth Publishers, New York, fourth edition, 2001.

[97] P. Morasso. Spatial control of arm movements. *Experimental Brain Research*, 42:223–227, 1981.

[98] L. Morency, A. Rahimi, and T. Darell. Adaptive view-based appearance model. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[99] P. Morissette, M. Ricard, and T. Gouin-Decarie. Joint visual attention and pointing in infancy: A longitudinal study of comprehension. *British Journal of Developmental Psychology*, 13:163–177, 1995.

[100] J. R. Movellan. An infomax controller for real time detection of social contingency. In *Proceedings of the 4th IEEE International Conference on Development and Learning*, 2005.

[101] J. R. Movellan and J. S. Watson. The development of gaze following as a bayesian systems identification problem. In *Proceedings of the 2nd IEEE International Conference on Development and Learning*, 2002.

[102] D. Muir and A. Slater. The scope and methods of developmental psychology. In A. Slater and G. Bremner, editors, *An introduction to developmental psychology*, pages 3–33. Blackwell Publishing, Oxford, 2003.

[103] C. M. Murphy and D. J. Messer. Mothers, infants and pointing: a study of a gesture. In H. R. Schaffer, editor, *Studies in mother-infant interaction*, pages 325–354. Academic Press, London, 1977.

[104] L. Murray and C. Trevarthen. Emotional regulation of interactions between two-month-olds and their mothers. In T. M. Field and N. Fox, editors, *Social perception in infants*, pages 101–125. Ablex, Norwood, N.J., 1985.

[105] F. A. Mussa-Ivaldi and E. Bizzi. Motor learning through the combination of primitives. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, 355:1755–1769, 2000.

[106] J. Nadel, I. Carchon, C. Kervella, D. Marcelli, and D. Reserbat-Plantey. Expectancies for social contingency in 2-month-olds. *Developmental Science*, 2(2):164–173, 1999.

[107] Y. Nagai. *Understanding the development of joint attention from a viewpoint of cognitive developmental robotics.* PhD thesis, Osaka University, 2004.

[108] Y. Nakamura and H. Hanafusa. Inverse kinematic solutions with singularity robustness for robot manipulator control. *ASME Journal of Dynamic Systems, Measurement and Control*, 108:163–171, 1986.

[109] J. Nakanishi, J. Morimoto, G. Endo, G. Cheng, S. Schaal, and M. Kawato. Learning from demonstration and adaptation of biped locomotion. *Robotics and Autonomous Systems*, 47:79–91, 2004.

[110] K. M. Newell, D. M. Scully, and P. V. McDonald. Task constraints and infant grip configurations. *Developmental Psychobiology*, 22(8):817–832, 1989.

[111] K. Gold P. Michel and B. Scassellati. Motion-based robotic self-recognition. In *Proceedings of 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2004.

[112] J. Park and I. W. Sandberg. Universal approximation using radial-basis-function networks. *Neural Computation*, 3(2):246–257, 1991.

[113] T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78(9):1481–1497, 1990.

[114] T. A. Poggio. A theory of how the brain might work. *Cold Spring Harbor Symposia on Quantitative Biology*, 55:899–910, 1990.

[115] A. Pouget and L. H. Snyder. Computational approaches to sensorimotor transformations. *Nature Neuroscience*, 3:1192–1198, 2000.

[116] D. J. Povinelli and D. R. Davis. Differences between chimpanzees (*Pan troglodytes*) and humans (*Homo sapiens*) in the resting state of the index finger. *Journal of Comparative Psychology*, 108:134–139, 1994.

[117] M. J. D. Powell. The theory of radial basis function approximation in 1990. In W. Light, editor, *Advances in numerical analysis Vol. II: Wavelets, subdivision algorithms, and radial basis functions*, pages 105–210. Oxford Science Publications, Oxford, 1992.

[118] P. C. Quinn, P. D. Eimas, and S. L. Rosenkrantz. Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants. *Perception*, 22:463–475, 1993.

[119] H. Ritter, T. Martinetz, and K. Schulten. *Neural computation and self-organizing maps: An introduction.* Addison-Wesley, New York, 1992.

[120] G. Rizzolatti, L. Fadiga, V. Gallese, and L. Fogassi. Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, 3:131–141, 1996.

[121] M. Scaife and J. S. Bruner. The capacity for joint visual attention in the infant. *Nature*, 253:265–266, 1975.

[122] B. Scassellati. Imitation and mechanisms of joint attention: A developmental structure for building social skills on a humanoid robot. In C. Nehaniv, editor, *Computation for metaphors, analogy and agents, Vol. 1562 of Springer lecture notes in artificial intelligence*, pages 176–195. Springer Verlag, Berlin, 1999.

[123] B. M. Scassellati. *Foundations for a theory of mind for a humanoid robot.* PhD thesis, Massachusetts Institute of Technology, 2001.

[124] R. Shadmehr, F. A. Mussa-Ivaldi, and E. Bizzi. Postural force fields of the human arm and their role in generating multijoint movements. *The Journal of Neuroscience*, 13(1):45–62, 1993.

[125] R. Shadmehr and S. P. Wise. Motor learning and memory for reaching and pointing. In M. S. Gazzaniga, editor, *The Cognitive Neurosciences III*, pages 511–524. MIT Press, Cambridge, M.A., 2004.

[126] R. Shadmehr and S. P. Wise. *The computational neurobiology of reaching and pointing*. MIT Press, Cambridge, M.A., 2005.

[127] L. G. Shapiro and G. C. Stockman. *Computer vision*. Prentice Hall, Upper Saddle River, N.J., 2001.

[128] W. Skrandies. The upper and lower visual field of man. In H. Autrum, editor, *Progress in sensory physiology*, volume 8, pages 162–189. Springer-Verlag, Berlin, 1987.

[129] J. F. Sorce, R. N. Emde, J. J. Campos, and M. D. Klinnert. Maternal emotional signaling: Its effects on the visual cliff behavior of 1-year-olds. *Developmental Psychology*, 21:195–200, 1985.

[130] G. Sun and B. Scassellati. Reaching through learned forward model. In *Proceedings of the 4th IEEE-RAS/RSJ International Conference on Humanoid Robots*, 2004.

[131] E. Thelen. Learning to walk: ecological demands and phylogenetic constraints. *Advances in Infancy*, 3:213–260, 1984.

[132] E. Thelen, D. Corbetta, and J. P. Spencer. Development of reaching during the first year: role of movement speed. *Journal of Experimental Psychology: Human Perception and Performance*, 22:1059–1076, 1996.

[133] E. Thelen, G. Schoner, C. Scheier, and L. B. Smith. The dynamics of embodiment: a field theory of infant perseverative reaching. *Behavioral and Brain Sciences*, 24(1):1–34, 2001.

[134] A. R. Tilley and Henry Dreyfuss Associates. *The measure of man and woman.* John Wiley & Sons, New York, 2001.

[135] C. Trevarthen. Communication and cooperation in early infancy. In M. Bullowa, editor, *Before speech: The beginnings of interpersonal communication*, pages 321–347. Cambridge University Press, Cambridge, 1979.

[136] J. Triesch, C. Teuscher, G. Deak, and E. Carlson. Following: why (not) learn it? *Developmental Science*, 9(2):125–147, 2006.

[137] Y. Uno, M. Kawato, and R. Suzuki. Formation and control of optimal trajectory in human multijoint arm movement: Minimum torque-change model. *Biological Cybernetics*, 61:89–101, 1989.

[138] V. N. Vapnik. *The nature of statistical learning theory.* Springer Verlag, New York, second edition, 1999.

[139] B. Vereijken, R. E. A. van Emmerik, H. T. A. Whiting, and K. M. Newell. Free(z)ing degrees of freedom in skill acquisition. *Journal of Motor Behavior*, 24:133–142, 1992.

[140] C. von Hofsten. Eye-hand coordination in the newborn. *Developmental Psychology*, 18:450–461, 1982.

[141] C. von Hofsten. Structuring of early reaching movements: a longitudianal study. *Journal of Motor Behavior*, 23(4):280–292, 1991.

[142] L. S. Vygotsky. Development of the higher mental functions. In A. N. Leontier, editor, *Psychological research in the U.S.S.R.* Progress Publishers, Moscow, 1966.

[143] J. S. Watson. Smiling, cooing and “the game”. *Merrill-Palmer Quarterly*, 18:323–339, 1972.

[144] J. S. Watson. Perception of contingency as a determinant of social responsiveness. In E. B. Thoman, editor, *The origins of social responsiveness*, pages 33–64. Erlbaum, New York, 1979.

[145] J. F. Werker. Becoming a native listener. *American Scientist*, 77:54–59, 1989.

[146] D. E. Whitney. Resolved motion rate control of manipulators and human prostheses. *IEEE Transactions on Man-Machine Systems*, 10(2):47–53, 1969.

[147] D. Wilkins. Why pointing with the index finger is not a universal (in sociocultural and semiotic terms. In S. Kita, editor, *Pointing: Where language, culture and cognition meet*, pages 171–216. Lawrence Erlbaum Associates, Mahwah, N.J., 2003.

[148] M. Williamson. *Robot arm control exploiting natural dynamics.* PhD thesis, Massachusetts Institute of Technology, 1999.

[149] J. M. Wolfe and G. Gancarz. Guided search 3.0: A model of visual search catches up with jay enoch 40 years later. In V. Lakshminarayanan, editor, *Basic and clinical applications of vision science*, pages 189–192. Kluwer Academic, Dordrecht, Netherlands, 1996.

[150] D. M. Wolpert, Z. Ghahramani, and M. I. Jordan. Perceptual distortion contributes to the curvature of human reaching movements. *Experimental Brain Research*, 98:153–156, 1994.

[151] K. Wynn. Addition and subtraction by human infants. *Nature*, 358:749–750, 1992.

[152] T. Yoshikawa. Manipulability of robotic mechanisms. *International Journal of Robotics Research*, 4(2):3–9, 1985.