DEVELOPMENTAL MODEL OF SENSORIMOTOR MAP ACQUISITION FOR A HUMANOID ROBOT

BY

LYDIA MAJURE

DISSERTATION
Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Electrical and Computer Engineering in the Graduate College of the University of Illinois at Urbana-Champaign, 2013

Urbana, Illinois

Doctoral Committee:
Professor Stephen Levinson, Chair
Associate Professor Todd Coleman
Professor Mark Hasegawa-Johnson
Professor Thomas Huang
Cognitive developmental robotics unites machine learning and neuroscience with the aim of creating robots which display the robustness and efficiency of human cognition. This document proposes a novel method for acquiring a visuomotor mapping for hand-eye coordination. This model is trained on the iCub humanoid robot and used for smooth control of reaching. Applications of this model to sensorimotor associative learning are examined. In addition, a derivation of the self-organizing map from neural dynamics is presented. This dissertation is part of an overarching goal of developing a platform-independent model of intelligent systems.
ACKNOWLEDGMENTS

My advisor, Dr. Stephen Levinson, has made this work possible. His broad perspective and depth of knowledge guided my approach to cognitive science. I have learned from him the value in taking intellectual risks. Furthermore, his efforts lead to the Language Acquisition and Robotics group’s award of the iCub humanoid robot, central to the research presented here.

The other members of my doctoral committee, Dr. Todd Coleman, Dr. Mark Hasegawa-Johnson, and Dr. Thomas Huang, have all been valuable contributors to this process. Tom Huang, as a distinguished leader of the computer vision community, wisely encouraged me to not get bogged down in vision problems when it was not the focus of my research. Todd Coleman has been a strong influence on my thinking about neural computation; many of my conversations with him are reflected in this work. In addition, he can be at least partially credited in my move toward neuroscience as an area of study. Mark Hasegawa-Johnson was my undergraduate thesis advisor, and was an early cheerleader for me when I was starting to consider graduate school. Throughout my time at the University of Illinois, Mark has been at my talks asking the most interesting and challenging questions.

Setting up and learning to use the iCub was a very challenging aspect of my experience in the Language Acquisition and Robotics group. I owe a debt of gratitude to the entirety of the global iCub community. I especially would like to thank Dr. Giorgio Metta, Dr. Paul Fitzpatrick, Dr. Lorenzo Natale, and Dr. Ugo Pattacini for their leadership in the iCub project and their infinite patience with us graduate students and our (sometimes dumb) questions.

My colleagues in the Language Acquisition and Robotics group have been instrumental in my success. Aaron Silver and Luke Wendt made great robot repairmen, fixing the mechanical difficulties we had on the way. Conversations with Alex Duda and Felix Wang on neuronal dynamics were the seed
which led to some of the analysis of the model presented here. Jacob Bryan and Onyeama Osuagwu joined the lab as I was finishing my degree, I appreciate their commitment to learning to use the iCub so that my code does not become entirely obsolete upon my departure. Foremost among my colleagues I would like to thank Logan Niehaus. Logan and I worked together to turn our iCub from a fancy sculpture into a usable robot. He served as a second set of eyes when I could not find the source of a bug, and did much of the drudgery of writing needed support code which was not included in the iCub repository. On top of it all, he is one of my closest friends; both his technical and personal support were central to my success.

During graduate school, I developed a motley crew of friends, many of whom were graduate students themselves. I could not have kept my sanity without the regular escape from our workload in the form of darts games, karaoke nights, and other displays of frivolity. There are far too many people to name here, but they certainly know who they are.

My dog, Anya, faithfully sat next to me while I wrote both my MS thesis and this dissertation. She is an exemplar of the phrase “(wo)man’s best friend”, and I am grateful to have had her with me during this journey. Besides this, she has provided me with perspective on the subject of cognition. She is a constant reminder to not let anthropocentric biases about intelligence cloud my judgment. I approach the issue of humanoid cognition with humility because I live with a “lower” animal which is in most ways superior to any robot or artificially intelligent system.

My parents have supported me in everything I have attempted. They taught me to value the life of the mind, to boldly dive into challenges, and to prioritize my own happiness over the judgments of others. Despite being actively discouraged by a number of public school teachers, they told me that I should pursue a technical field. During graduate school, funding can sometimes be ephemeral, disappearing with a moment’s notice. They were instrumental in keeping a roof over my head when such unexpected situations happened. In today’s world, intellectualism is often dismissed as elitist. I am extremely grateful to my parents for teaching me the opposite: money can buy a lot of things, but it cannot buy intelligence. They raised me to believe that valuing learning and careful thought is an egalitarian virtue, one of the few avenues through which us regular folk can influence our world.

Finally, I would like to thank my husband, Mustafa Mir. He is my best
friend and my partner in all things. As we are in the same program, he has encouraged me in a way that only one who truly understands what someone is going through can. He has listened to endless rambles about small technical problems and rants about the many people and things I have disliked. At the times when I have been most cynical, his passion for science has reminded me what I was doing here to begin with. Certainly, there are many good outcomes from my decision to pursue my doctorate at the University of Illinois. I feel confident saying, however, that having met the love of my life here was the best.
# TABLE OF CONTENTS

**LIST OF ABBREVIATIONS** .................................................. viii

**CHAPTER 1  INTRODUCTION** ........................................... 1  
  1.1 Defining Intelligence ............................................ 1  
  1.2 The Internal Representation of Language ....................... 2  
  1.3 Cognitive Developmental Robotics ............................... 3  
  1.4 Developmental Motor Learning ................................. 4  

**CHAPTER 2  PRELIMINARIES** ......................................... 6
  2.1 Language Acquisition and Robotics Group ..................... 6  
    2.1.1 Navigation and Tracking ............................ 6  
    2.1.2 Emergent Semantics ............................... 7  
    2.1.3 Syntax Learning ................................... 8  
  2.2 Machine Learning ............................................. 8  
    2.2.1 Biologically Inspired Models ....................... 9  
    2.2.2 Topological Maps ................................ 12  
    2.2.3 Hidden Markov Models ............................ 12  
    2.2.4 Computer Vision .................................. 14  
    2.2.5 Learning Dynamics ................................ 14  
  2.3 Human Cognition ............................................... 15  
    2.3.1 Neural Dynamics .................................... 15  
    2.3.2 Motor Development ................................ 15  
    2.3.3 Vision .................................................. 17  
    2.3.4 Sensorimotor Integration and Language ............ 17  
  2.4 Cognitive Robotics ........................................... 18  
    2.4.1 Motor Learning and Spatial Awareness ............ 18  
    2.4.2 Imitation ............................................. 19  
    2.4.3 Multimodal Associative Memory .................. 19  
  2.5 The iCub Robot .............................................. 20  
  2.6 Open Questions on Cognition .................................. 21  

**CHAPTER 3  AUTONOMOUS LEARNING OF HAND-EYE COORDINATION** ................................. 23  
  3.1 Training Procedure ........................................... 23  
    3.1.1 Motor Babbling ...................................... 23
3.1.2 Hand Localization .................................. 24
3.2 Visuomotor Mapping ................................. 26
  3.2.1 Motor SOM for Kinematic Modeling ............ 26
  3.2.2 Model Performance ............................... 28
3.3 Dynamic Reaching .................................. 33
3.4 Discussion ........................................... 34

CHAPTER 4 MOTOR PRIMITIVES FOR ASSOCIATIVE LEARN-
  ING .................................................. 36
  4.1 Features of Self-Organizing Motor Model .......... 36
    4.1.1 Implicit Continuity of Representation .......... 36
    4.1.2 Model Decoupling ................................ 37
  4.2 Grasp Acquisition .................................. 37

CHAPTER 5 TOPOLOGICAL MAPPINGS FROM NEURONAL
  DYNAMICS ........................................... 41
  5.1 Scale-Free Networks of Neurons .................... 41
  5.2 Plasticity Models .................................. 42
  5.3 Topology-Learning Neural Network ................ 42
  5.4 Discussion ....................................... 46

CHAPTER 6 CONCLUSION ................................. 47
  6.1 Mapping between Sensory Reference Frames ........ 47
  6.2 SOM as Model of Topological Learning ............ 48
  6.3 Associative Memory ................................ 48
  6.4 Future Work ...................................... 49
  6.5 Platform-Independent Model of Intelligence .. 49

REFERENCES ............................................. 51
LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAR</td>
<td>Language Acquisition and Robotics</td>
</tr>
<tr>
<td>HMM</td>
<td>Hidden Markov Model</td>
</tr>
<tr>
<td>CHMM</td>
<td>Cascaded Hidden Markov Model</td>
</tr>
<tr>
<td>RMLE</td>
<td>Recursive Maximum-Likelihood Estimation</td>
</tr>
<tr>
<td>SOM</td>
<td>Self-Organizing Map</td>
</tr>
<tr>
<td>RBF</td>
<td>Radial Basis Function</td>
</tr>
<tr>
<td>STDP</td>
<td>Spike-Timing Dependent Plasticity</td>
</tr>
</tbody>
</table>
CHAPTER 1
INTRODUCTION

1.1 Defining Intelligence

In popular discourse, “intelligence” is used to refer to many phenomena. It is applied to anything from highly skilled humans, to adaptive algorithms, to simply data (for example, “Central Intelligence Agency”). Of course, this is a linguistically acceptable situation. Words are defined by the people who use them, and by the brain processes that occur upon speaking and hearing them. This subject is to be revisited in Section 1.2, but for now we must agree upon what this word means in a scientific context. An intelligent system is one which can both build a meaningful representation of its input, and behave in such a way as to acquire more data or otherwise benefit the system. Humans are, of course, the most prominent and impressive example of this system property, but intelligence in the most general sense is displayed by all living organisms. Intelligence is an input, an output, and a model. Because of limited storage space, so to speak, all organisms form an incomplete model of their environment. The incoming data which is not behaviorally useful is discarded as it comes in, or even during the evolutionary process, as blind cave animals show.

In information theoretic terms, the brain’s memory structures perform lossy compression on the sensory input stream. This process is optimized for minimizing the size of the representation without compromising behavioral goals. This lossy compression serves an important function beyond saving space; it simplifies learning and decision making. The loss of information in the model leads to generalization, which allows for flexible behavior in novel situations. The concept of learning as compression is reflected in many basic pattern recognition techniques, such as Gaussian mixture models and principal component analysis, where simpler models are learned which ap-
proximate the data. Even the theory behind algorithm-independent machine
learning reflects this. The “no free lunch” theorem is a consequence of the
inherently lossy nature of learning. Despite the prevalence of this perspec-
tive in the field of pattern recognition and machine learning, it is relatively
uncommon to encounter such a perspective in cognitive science. However, it
is highly instructive to consider brains in the context of the mathematical
fundamentals of information theory.

1.2 The Internal Representation of Language

The human skill seen as most exemplary of our intelligence is language. This
is not only because of its uniqueness to humans (as far as we know), but
also because it is the foundation of several complex behaviors. It was once
thought that human language processing was entirely based on the acous-
tic signals which comprise words. However, bottlenecks in purely acoustic
speech processing technology suggest otherwise. An even stronger counter-
argument to this idea is evidence from brain imaging, which demonstrates
that language activates large swaths of the brain beyond the auditory cortex.
Language seems to be an inherently multi-sensory phenomenon. The reason
this should be so can be understood from the very purpose of communica-
tion: generating a shared mental state between speaker and listener. When
linguists refer to semantics, they are discussing the relationship between the
word as an acoustic symbol and the word as a distributed cortical activity
pattern. One of the simplest examples can be taken from words for colors.
The word “blue” is a symbol which refers to the activation of certain types of
color cells in the visual cortex, regardless of position in the visual field. Sim-
ilarly, the word “up” refers to a certain region of the body centered spatial
map, regardless of the objects recognized as being in that area. Certainly,
most words represent a far more complicated activity pattern than these;
even these two examples can also be used to refer to more than just simple
peripheral sensory phenomena.

Because semantics are defined by cross-sensory associations, traditional
speech processing methods are inherently limited. Most automatic speech
recognition systems break catastrophically in the presence of even moderate
noise. On the other hand, humans can have conversation in the noisiest of
environments, even sometimes making do when it is practically impossible
to hear. This is because the multi-sensory, distributed nature of language
is more than just a representation, it serves as a powerful error correction
system. Acoustically similar words are disambiguated by accounting for the
multimodal context in which they are being heard. While standard speech
recognition computes the maximum likelihood of a word given the acous-
tic signal, the brain computes the maximum likelihood of a word given the
entire sensory experience and all world knowledge. In Searle’s famous “Chi-
inese Room” thought experiment, a non-sinophone sits in a room with an
extremely thorough codebook on the Chinese language [1]. It is so thor-
ough, in fact, that any slip of paper passed to the operator in Chinese will
be matched with a correct, sensible response and passed back out. Searle’s
critique of strong artificial intelligence in this story is that there is nowhere
in the system where true intelligence resides. If such a codebook could ac-
tually be constructed, which is already an impossible task, it is true that it
would not be equivalent to a human mind. However, this does not disprove
the possibility of implementing an artificial mind, it merely points out the
importance of embodied experience of the world to language and cognition.
The Chinese Room is not intelligent because it has no idea what the words
really mean.

1.3 Cognitive Developmental Robotics

Since the essence of intelligence is building a model of the world, it is futile
to implement artificial intelligence with no sensory inputs. In addition, a
sufficiently nuanced model of the environment to behave flexibly cannot be
hard coded into a system, it must be acquired via interaction with the en-
vironment. The field of cognitive developmental robotics results from these
assertions. The idea that true intelligence requires embodiment is not new, it
dates back to Turing in 1950, in his seminal paper on artificial intelligence [2].
However, the development of robotics technology and computers sufficiently
powerful to implement this idea is quite recent. In the same era, the field
of cybernetics was proposed by Wiener [3]. Cybernetics proposed that both
machines and biological organisms could be understood through the lenses
of control theory and probability theory. Cognitive developmental robotics
is the descendant of both of these intellectual leaps.

The goal of this research is build an robot of intelligence comparable to that of a human, not to literally build a brain. In many respects, such as chemical signaling, the implementation details of the brain are not particularly important to understanding cognition. However, there are certain properties of human cognition which are important to any robotic implementation. Any intelligent agent must be able to both acquire the world model and make behavioral decisions from it simultaneously. In addition to an array of external senses, the agent must have proprioception, an awareness of its own body state. Behaviorally unimportant information needs to be discarded, and novel stimuli need to be emphasized. Noise and variation in observational conditions must minimally affect ability to recognize objects. An intelligent agent must be able to make decisions by predicting the outcome of actions based on prior experience. These behavioral aspects of human cognition are likely to be present in any intelligent system.

1.4 Developmental Motor Learning

Humans are born barely even able to move voluntarily, but quickly become adept at navigating their environment and manipulating objects. Our very limited motor repertoire at birth consists of only primitive reflexes, such as suckling and grasping. However, infants quickly learn to look at objects of interest, and soon after learn to reach for them. With time, they learn to make purposeful actions, both by trial and error and by imitation of adults. By adulthood, motor skills are impressively advanced. Even by early childhood, we can walk on uneven ground without looking down. Activities from surgery to martial arts demonstrate the ability of humans to learn highly precise yet adaptive motor skills. This is in contrast to the mainstream robotics research of the last 50 years. Most robotic systems perform perfectly under a limited set of specified conditions and tasks, but fail spectacularly outside of that operation domain. Much like the artificial intelligence community ignored robotics, the robotics community paid little attention to machine learning and cognitive science until recently, focusing on classical control systems instead.

The motor system, for evolutionary reasons, is a hybrid of many types of
structures and algorithms. At the lowest level, in the spinal cord, there are simple feedback control systems; these are a very ancient inheritance from the earliest animal nervous systems. Further up the pathway, the cerebellum is thought to be a primarily feed-forward system which calculates a smooth trajectory for cortically initiated movements. Finally, many processes take place in the cortical motor pathways. Proprioceptive information is integrated for behavioral decisions in topological maps of the body in the primary motor and somatosensory cortices. Mappings between the visual field, body pose, and sound are assembled in the parietal lobe. This research focuses primarily on the cortical aspects of motor control. The prerequisite for imitation is hand-eye coordination, the robot must be able to move its limbs to the appropriate point in space. To learn from demonstrated behavior, the humanoid robot must be able to map the observed body to its own. Finally, for meaningful actions to be learned, the motor representations need to be accessible to the language centers, such that demonstrations can be labeled. This study will address these primary issues in order to advance the state of developmental motor learning in humanoid robots.
CHAPTER 2
PRELIMINARIES

This chapter is intended to discuss the body of scientific and engineering research relevant to motor learning, imitation, and language acquisition by robots. First, the history of the Language Acquisition and Robotics group at the University of Illinois at Urbana-Champaign will be reviewed. In the second section, machine learning techniques applicable to this research will be discussed. In the third section, studies on human cognition and neuroscience will be covered. The fourth section will review the state of the art in cognitive robotics research. A fifth section describes the iCub humanoid robot, the platform for the experiments described in Chapters 3 and 4. Finally, open questions in the field will be discussed.

2.1 Language Acquisition and Robotics Group

The Language Acquisition and Robotics Group (LAR) at the University of Illinois at Urbana-Champaign is dedicated to the goal of developing intelligent robots capable of learning language via interaction with the world. The methods used are rooted in the framework of cognitive developmental robotics. Until recently, experiments have been done on non-humanoid multi-sensory mobile robots. These robots featured three wheels, binocular vision, sound localization, and collision detection whiskers. Presently, the LAR group uses an iCub humanoid robot, discussed in Section 2.5.

2.1.1 Navigation and Tracking

A central problem in adaptive mobile robotics is learning navigation. Reinforcement learning is a very popular family of algorithms for this task. One online reinforcement learning technique, Q-learning, was adapted for
the LAR robots. The method was dubbed propagated Q-learning, or PQ-learning, because it learned adjacent states to the winner at each time step [4]. This improved efficiency and robustness, allowing the robot to successfully navigate to locations not previously visited. To be able to learn about objects, the robot needs to be able to track them from frame to frame as it moves through the environment. Toward this end, a technique called multiview object recognition was used [5]. Using a shape-based visual feature space, objects were parameterized as a mixture of Gaussians. If there was sufficient ambiguity, the robot would automatically move to obtain a new viewpoint of the object.

2.1.2 Emergent Semantics

The core philosophy of the LAR group is that semantics are rooted in association between different sensory streams. To accomplish autonomous language acquisition, a model is needed which can learn cross-sensory correlations in an unsupervised manner. The Cascaded Hidden Markov Model (CHMM) was introduced to address this problem [6]. In a CHMM, multiple input observations are learned by a Hidden Markov Model (HMM); in turn, a model of these input states is learned by a top-level HMM. The functional compositionality of this model allows learning to occur at multiple levels and can represent multimodal input. Figure 2.1 illustrates this model in the context of audiovisual association.
The Recursive Maximum-Likelihood Estimation (RMLE) algorithm was used to implement online learning of the model. HMM will be treated in more detail in Section 2.2.3. Using online training of a CHMM, the robots were able to learn words for objects in their environment from human instruction, demonstrating the representational power of this type of model. In addition, CHMM can be used as a generative model, allowing the robot to reproduce learned words upon seeing a recognized object. In order to implement autonomous exploration, robot decisions were generated by giving the robot a Finite State Machine controller, including human interaction, learning, and object interaction behaviors. By learning concepts with no a priori information about the world, this model captures the way that semantics emerge from real-world experience. More recently, this framework has been used to learn action-word pairs for the iCub humanoid robot [7].

2.1.3 Syntax Learning

Building on the work modeling semantics as multimodal associative memory, a study was conducted in which robots learned simple syntactic structure from a semantic bootstrapping process [8]. In addition to shapes of objects and speech, the associative memory had representations of object movement and robot action. A discrete HMM was used to learn the grammar, with three hidden states learned as lexical categories and four observations corresponding to relevant modality. Figure 2.2 shows the relationship between models. The robot was presented with subject-verb sentences which described the action it had chosen to perform. Using RMLE to learn the model online, the three states converged to representing noun, verb, and delimiter, with the transition probabilities representing the production rule. This result demonstrated the possibility of learning grammar without a priori knowledge or specialized syntactic learning structures.

2.2 Machine Learning

Machine learning is a broad term which encompasses much of modern computational science. This study, however, is primarily concerned with connectionist methods. Common features of connectionist methods include simi-
larity to neural processes, probabilistic models, and emergence of complex system behavior from simple units. Within this paradigm in machine learning, the types of algorithms most relevant to the research presented here are those that are both unsupervised and online.

2.2.1 Biologically Inspired Models

Classical Neural Networks

The earliest neural network introduced was the perceptron in 1958 [9]. The perceptron is a one-layer, linear feed-forward network; it is essentially a learned linear discriminant. Because a perceptron can only represent hyperplane boundaries in the input space, its applicability is limited. Multi-layer networks with nonlinearities are much more expressive, and have been demonstrated to be able to learn arbitrary functions when trained using the back-propagation algorithm [10]. Back-propagation networks have become popular in the decades since their introduction due to their versatility and claim to biological inspiration. However, there is little evidence that back-propagation networks function in a similar way to any brain structure. Furthermore, they are necessarily supervised learning methods, which are of limited use in autonomous mental development studies.
Unsupervised Neural Networks

The unsupervised branch of neural network research is more directly rooted in neuronal dynamics. Hebbian learning is a description of neuronal plasticity which attempts to account for how associations are formed in the brain; colloquially, it is phrased as “fire together, wire together” [11]. The formula stated by Hebb in his original work does not lead to stable networks, but it has influenced a number of unsupervised learning methods. Oja’s rule is a stable generalization of Hebbian learning which has demonstrated ability to learn principal components of the input space [12]. A form of Hebb-Oja learning has also been demonstrated to learn principal subspaces for dimensional scaling [13]. In spite of the difficulties implementing Hebbian learning as a machine learning method, the principle of learning being triggered due to activity at both input and output has remained a guiding principle of many biologically inspired models.

Hopfield and Boltzmann networks are similar models which demonstrate pattern completion and autoassociation [14], [15]. In both cases, the network functions by learning equilibrium states corresponding to the statistics of the data presented; the difference is that the Hopfield network is equivalent to a classical deterministic dynamical system, and the Boltzmann network is stochastic. Boltzmann chains, a trellis-like network structure with both hidden and visible units, are in fact equivalent to the HMM [16]. Botzmann networks are extremely computationally intensive, but some modifications have been proposed which improve efficiency, such as one which trains a hierarchical network by beginning with a greedy upward training network, then refining weights with a “sleep-wake” training cycle [17].

Supervised Recurrent Networks

Recurrent networks are capable of representing temporal structure in the input data, in contrast to previously mentioned networks. Continuous-time recurrent neural networks are one particularly interesting variety which have been shown to be universal dynamical approximators which can model attractors, limit cycles, and even chaotic trajectories [18]. This property has led to this type of network being applied to nonlinear model predictive control [19]. Despite the theoretical representational power of many varieties of
recurrent network, the whole genre is plagued by difficulties with learning rules and convergence. Reservoir computing is a type of network intended to address these issues. In general, these networks contain a large, randomly connected reservoir of neurons with fixed weights. The only part of the network which is “visible” and trainable are the input and output units, and their weights to the reservoir neurons [20]. This network structure reduces the computational load of using recurrent neural networks for applications such as robotics.

Cortical Models

Better understanding of the brain at a systems level has led to several techniques which mimic the high level computations of the cortex. One region of the brain which is particularly well understood is the primary visual cortex, providing a tempting subject for cortical modeling. Based on the observation that the equivalent of 2D Gabor filters can be found in this area as edge detectors, neural networks have been implemented which learn this transform [21]. Another approach created edge detecting networks using evolutionary reinforcement learning [22]. Due to the complexity of the motor system, it is less studied how to apply highly detailed cortical models to motion. However, simple arm control using a realistic neural microcircuit has been demonstrated [23].

Another type of model which can be called cortical in a quite different way is the Hierarchical Temporal Model [24]. It is similar to the CHMM in that it accommodates temporal nesting of sequences and widening receptive fields. Although this method mimics the connectivity structure of cortical microcolumns, the algorithms used at nodes in the network are classic pattern recognition techniques. This type of model is especially adept at vision problems, but has also been used for unsupervised phoneme acquisition to some success [25]. A similar model, Hierarchical Quilted SOM, is a nested structure of coupled recurrent and standard SOM, with widening receptive fields at higher layers [26].
2.2.2 Topological Maps

Self-organizing topological maps are also a biologically inspired method, but due to their interesting properties and high relevance to the research presented here, they merit their own discussion. The self-organizing map (SOM) was first described by Kohonen, hence the alternate name, Kohonen Map [27]. The original model consists of units arranged in a grid of any dimensionality, although one to three is typical. For unit $i$ and input vector $x$, the activation $y_i$ is calculated as

$$y_i = w_i^T x \quad (2.1)$$

For winner $i^*$ and learning rate $\eta$ the weights are updated by

$$\Delta w_i = \eta \Lambda(|i - i^*|)(x - w_i) \quad (2.2)$$

where $\Lambda$ is a neighborhood function on the grid, such as Gaussian or Mexican Hat, which is centered on the winning neuron $i^*$. The neighborhood function creates the topological representation of the SOM, as it forces adjacent neurons to represent nearby regions of the input space. Modifications to this have been proposed which prevent formation of dead units and speed convergence, such as habituation to repeated inputs [28].

Variations on the Kohonen map exist which optimize different criteria. The kernel-based maximum entropy learning rule forms an equiprobabilistic topographic map of the input using overlapping, variable-radius RBF kernels [29]. This approach maximizes neuron entropy instead of minimizing mapping distortion, a property which seems to be present in many cortical topological maps. Another technique, neural gas, forms the map in a different manner; populating the input domain with neurons and learning lateral connections between them as the topology [30].

2.2.3 Hidden Markov Models

Hidden Markov Models (HMM) have long been the standard method for automatic speech recognition. Their representational capabilities, however, can be extended to a wide range of time sequence recognition problems. The simplest version of the HMM, that with discrete observations, is best for illustrating the concept. The HMM can be characterized by three parameters:
π, the initial state distribution, A, the state transition probability matrix, and B, the observation probability matrix. Estimation of these parameters is performed by the Baum-Welch algorithm [31]. The forward probability, α$_t$(i), is the probability that the system is in state i at time t, given all past observations. For N states, M output symbols, observation sequence ($O_1...O_T$), and assuming α$_1$(i) = π$_i$b$_i$(O$_1$), it is computed as

$$\alpha_{t+1}(j) = \sum_{i=1}^{N} \alpha_t(i)a_{ij}b_j(O_{t+1}), \quad 1 \leq t \leq T - 1$$

The backward probability, β$_t$(j), is the probability of being in state j at time t, given all future observations. Assuming β$_T$(j) = 1,

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij}b_j(O_{t+1})\beta_{t+1}(j), \quad T - 1 \geq t \geq 1$$

The probability of the observation given the model can be computed simply as

$$P = \sum_{i=1}^{N} \alpha_T(i)$$

The reestimation formulas use these quantities to maximize the probability of the observation given the model as follows:

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)}$$

$$\bar{b}_{jk} = \frac{\sum_{t \geq O_t=v_k} \alpha_t(j)\beta_t(j)}{\sum_{t=1}^{T} \alpha_t(j)\beta_t(j)}$$

$$\bar{\pi}_i = \frac{1}{P} \alpha_1(i)\beta_1(i)$$

This procedure is guaranteed to increase P on each iteration until a local maximum is reached. HMM have been demonstrated to learn latent structure from a text. The model can discover patterns in English spelling from text, such as learning zero probability transitions between letters which never occur in sequence [32]. It has also been shown to distinguish broad categories of phonemes when given unlabeled speech [33]. The Baum-Welch algorithm requires complete knowledge of the training sequence, making it unsuitable for developmental learning approaches. The RMLE algorithm provides a
method for learning the parameters of the HMM online during training [6].

2.2.4 Computer Vision

Edge detection is essential for almost all vision tasks. Many algorithms exist to address this, but most work by finding the gradient in a local neighborhood, and differ only in the details. The most popular edge detection method is the Canny algorithm, which consists of convolving the image with a function which is approximately the first derivative of a Gaussian [34]. Banks of Gabor filters are another common choice for this filter’s similarity to the responses of the primary visual cortex [35]. A more complicated, biologically inspired method is nonclassical receptive field inhibition, which first calculates edges locally then inhibits edges depending on their neighboring values [36].

Motion tracking techniques divide into two main types: those which find correspondences between frames to locate the object of interest, and those which develop a model of object motion. The former is done primarily by feature matching such as edges or color. One method to detect moving objects is an optical flow algorithm [37]. The most common example of the latter is the Kalman filter, a linear predictive filter [38]. Particle filtering is a more general method which is not restricted to linear motion [39].

2.2.5 Learning Dynamics

The learning of dynamical systems is of great interest to robotics. While learning linear dynamics is a relatively easy problem, most problems in cognitive robotics require nonlinear dynamics. One method proposed for estimating such systems uses the expectation-maximization algorithm to fit the dynamics to radial basis functions (RBF) [40]. Another method, inspired by the HMM, uses a hidden and observed state variables structure and estimates the dependencies between them [41]. Evolved dynamical neural networks have also been demonstrated as a learning method [42]. Robotics results related to this research will be discussed in Section 2.4.
2.3 Human Cognition

2.3.1 Neural Dynamics

The Hodgkin-Huxley model is the most thorough description of the spiking behavior of individual neurons [43]. However, many models exist which do not perfectly reproduce observed spiking behavior but significantly reduce the computational load of simulating them [44]. Plasticity, change in structure and connectivity due to activity, is the central mechanism of learning in the brain. Spike-timing dependent plasticity is especially of interest due to its prevalence in cortical regions [45]. The STDP model most predictive of observed plasticity is BCM, however, much simpler models can be used to generate observed learning patterns [46]. Models combining Hebbian principles and STDP have succeeded in demonstrating the formation of stable learning networks [47]. Networks of spiking neurons undergoing STDP develop receptive fields and differentiate connectivity patterns based on the relative strength of spatial or temporal correlation in the input [48].

Neural dynamics are likely the basis for representation in the brain. Neural networks formed through a combination of synaptic and structural plasticity can serve as associative memories [49]. Certain types of subnetworks of spiking neurons have been shown to synchronize under stimulation, allowing memory to be stored as attractors in the phase space [50]. Structural features of neural networks also contribute to their representational power. The cortex is arranged into microcolumns which form local neural circuits [51]. There is significant evidence for computational similarity between microcolumns across all regions of the cortex [52]. Simulations show that the laminar structure of cortical microcolumns contributes to higher performance over non-laminar neural circuits of the same size [53]. The cerebellum, distinct from the neocortex in its structures, has been postulated to operate like a liquid state machine, a type of reservoir network [54].

2.3.2 Motor Development

Motor learning and control involves multiple nervous system structures. Voluntary motion is initiated and planned in the motor cortex, but the smooth trajectory is calculated in the cerebellum [55]. The classical viewpoint is that
the cerebellum and basal ganglia, unlike the cortex, are structures highly specialized to motion. However, research which demonstrates their involvement in non-motor functions has led to the hypothesis that each of the cortex, basal ganglia, and cerebellum are in fact specialized for different types of learning: unsupervised, reinforcement, and supervised, respectively [56].

Babbling and Reflexes

Infants are born with no motor skills save primitive reflexes such as grasping. These disappear during child development, but probably contribute to bootstrapping the motor system first [57]. Early random motor exploration guides development of hand-eye coordination and voluntary motion [58]. This process is often referred to as “motor babbling”, in analogy to the similar production of random sounds at the beginning of language development. These two processes of reflexive motion and random motion provide the training data the infant needs for its motor system, so to speak. For example, the tonic neck reflex assists with keeping the child’s hand in view for learning coordination, and the grasp reflex is a base movement which forms the basis of a later voluntary class of motions [57].

Motor Primitives

Although vertebrates control motion primarily from the brain, the spinal cord is capable of generating movements on its own [59]. These motions are often referred to as motor primitives or central pattern generators, and are thought to form the building blocks of voluntary motion. Motion in vertebrates is produced by superimposing motor basis fields [60]. It is possible that early reflexes are subsumed into primitives instead of entirely disappearing [61]. Studies on human arm motion have shown behavioral evidence for use of primitives when planning motion [62]. Humans fixate on the hand when being demonstrated an arm motion, leading to the suggestion that the whole arm motor signals needed to replicated the motion are inferred from an existing set of primitives [63].
Grasp Acquisition

Grasp acquisition begins separately from visuomotor coordination, early prehensile actions occur in the absence of objects to be grasped [64]. As development proceeds, reaching, grasping, and gaze become coupled in behavior [65]. Visual feedback plays an important role in grasp learning for different object shapes, hands assume an appropriate pose for grasping during the reach for the object [66]. Encoding of the relative position of hand, gaze, and target has been observed in premotor areas of the parietal cortex [67].

2.3.3 Vision

Vision is the primary way humans obtain information about their environment. The visual pathway consists of several regions which display different computational specializations. The primary visual cortex (V1) is a retinotopic map of the visual field which contains edge orientation selective cells [68]. Motion-selective cells can be found in both V1 and Middle Temporal (MT), but cells in V1 have smaller receptive fields which respond to visual component motion, whereas cells in MT respond more to pattern or object motion [69]. The Inferotemporal (IT) cortex shows specialization for shapes and objects [70]. In general, a tendency from more simple representations to more complex can be seen as one moves up the visual pathway. There is evidence for a division of labor between two main visual pathways. The ventral path is thought to be primarily related to object recognition and linguistic identification. The dorsal path is thought to be used for object location [71].

2.3.4 Sensorimotor Integration and Language

Sensorimotor integration is essential for behavior. Functional brain imaging studies indicate that representations of semantic concepts are distributed throughout the cortex [72]. For motor activities and navigation, topological maps from different reference frames must be integrated, such as the retinotopic map of the visual system and the somatotopic map of the proprioceptive system. This process seems to happen primarily in the parietal cortex, generating the information needed for coherent behavior [73].

Humans are learning from the moment of birth, and models of the world
are being formed significantly before linguistic labels are attached to the features. The earliest phase in language learning is a fine-tuning of phoneme differentiation to the native language [74]. Vocal babbling occurs next, allowing the infant to discover how to reproduce sounds it has heard. Full language acquisition occurs when the child associates strings of phonemes with semantic concepts. The storage of semantic concepts themselves does not require any special structure, as they are stored in the same areas used when perceiving the object [72]. This distributed representation extends to motor representation, with evidence that “action” words trigger activity in motor cortical areas [75].

2.4 Cognitive Robotics

2.4.1 Motor Learning and Spatial Awareness

An important challenge in robotics is learning hand-eye coordination and reaching. Knowing the exact kinematics is, of course, one method for solving this problem, but only learned methods are of interest here. One study designed a reservoir network which learned the forward and inverse kinematics of a simulated humanoid robot arm [76]. Another used SOM for a visuomotor mapping between end effector and visual position, using the gain and the noise of the controller to switch between reaching and random exploration [77]. Neural networks have also been used to learn both open and closed loop arm controllers [78]. While it is common to give a marker for the location of the hand, it is not necessary, as was demonstrated by a robot discovering its own hand via correlation between motor commands and visual motion [79].

Reaching is not the only type of movement which merits study. In the method of Stylistic Dynamic Movement Primitives, the speed and deviation of a movement are adjustable parameters in the learned control system [80]. It is also possible to learn shape primitives from critical points on a trajectory such that the robot is able to reproduce the trained shapes by drawing [81]. An interesting study addressing both speech and motor babbling simulated a human mouth and the sounds produced by it, then trained a SOM to associate the sound with the motor commands [82]. An evolutionary approach to motor learning was implemented using genetic algorithms
to develop continuous-time recurrent neural network controllers [83]. A similar approach can be used for finding optimal dynamical systems for learning trajectories [84].

2.4.2 Imitation

Imitation is a fairly new area of interest in robotics, brought about by increasingly humanoid robots capable of performing human-like actions. Many approaches exist, varying in algorithm and their degree of a priori knowledge. One study presented the technique of Incremental Local Online Gaussian Mixture Regression for imitating tasks. An interesting component of this experiment was the robot learning the appropriate spatial reference frame (self, object, or hand centered) for a given task [85]. Bayesian belief nets have been demonstrated for a non-humanoid robot acquiring primitive motions from human demonstration [86]. Continuous-time recurrent neural networks have found use in this goal, as well, due to their dynamical properties. Networks of fast and slow neurons have been demonstrated to self-organize into a functional hierarchy for complex motions [87]. HMM, because of their use in sequence recognition, have also been applied to imitation. Hierarchies of HMM have been applied to learning sequences of actions from motion capture of humans [88]. HMM have also been used in conjunction with impedance control to implement learning of primitives from direct kinesthetic instruction [89]. Estimation of nonlinear dynamics has been used for robot imitation of provided human motion capture data [90]. Gaussian Mixture Models have been used as an approach to incremental learning of gestures by imitation [91].

2.4.3 Multimodal Associative Memory

Sensorimotor integration is essential to the goal of understanding cognition. An important first step is presenting all sensory information in a coherent reference frame for action, this was implemented in the iCub robot for bottom-up attention [92]. Some approaches to language, while recognizing the importance of multimodal information, do not allow for true emergence because the concepts which the robot is to learn are preset [93]. Probabilistic
graphical models are another approach, demonstrated for learning of object properties on a humanoid [94]. Feed-forward neural networks with inputs from multiple modalities have been explored for simple non-humanoid robot associative memory [95]. Neural gas networks have been shown to simultaneously learn names for multiple attribute categories of objects, such as shapes and colors [96]. CHMM have been demonstrated as powerful models of associative memory. In one study which was done only with a camera and microphones, not on a robot, a network of HMM acquired the phonemes of Japanese, used them to learn words for presented objects, and to acquire a grammar [97]. This is similar to the experiments of the LAR group [6, 8, 7]. One experiment used SOM-based associative memory to implement distinct “what” and “where” pathways in the network, as found in the brain [98]. Associative memory frameworks have also been applied to grasping, to learn haptic and visual associations [99].

2.5 The iCub Robot

The iCub is a three-and-a-half-foot tall, 48-pound humanoid robot with 53 degrees of freedom in its joints. Figure 2.3 shows the iCub. It is one of the most advanced humanoids to be designed to date. It was designed specifically for studying cognitive processes in humans and robots. The software and hardware is entirely open source under the GPL. The hands have 9 degrees of freedom and are entirely tendon-driven. Each arm has 7 degrees of freedom. The iCub has a rich array of sensory information available to it. The eyes are two cameras with a 30 Hz frame rate. They have 3 degrees of freedom of motion defined in terms of the head-centered polar fixation. The ears are microphones with external pinnae for sound localization. Proprioceptively, there are joint position, speed, and torque readings available, as well as a gyroscope in the head. Low level joint control is done by DSP control boards. There is a PC104 in the robot head which synchronizes and reformats sensorimotor information from the robot. The robot is connected by an ethernet and power umbilical cord to a cluster of computers.

The backbone of the iCub software is Yarp. Yarp is a robotics communication library which allows a great degree of modularity and platform independence. The main features are a port class which is agnostic to the
communications protocol and data type, and a device class which wraps hardware drivers. Robotics has long been plagued with ad hoc software; Yarp is meant to address this issue. In addition to Yarp, there is an iCub-specific repository which includes motor interfaces and a robot simulator. A detailed description of the hardware and software of the iCub can be found in [100].

2.6 Open Questions on Cognition

Studies in motor learning and imitation using the concept of motor primitives from neurophysiology have shown promising preliminary results. Likewise, studies in multimodal associative memory for semantic and syntactic learning have demonstrated developmentally grounded language understanding in robots. These studies need to be continued and expanded upon to advance understanding of cognition. As the basics of artificial intelligence begin to be understood, deeper theoretical questions start to arise. One that is already becoming important is the level of abstraction which is acceptable without compromising the integrity of the cognitive model. It is widely accepted that first-order logic machines are far too much abstraction, and that modeling at the molecular level is far too little. However, there is abundant disagreement
as to whether abstraction should occur at the neuron, microcolumn, or functional region level. As brain research advances, it will become possible to model at any of these levels. However, an optimal level of modeling will need to be developed for cognitive robotics applications. A related question is how to describe the processes of cognition in a mathematical way, independent of platform. These answers will become clearer in time with a combination of evidence from robotic studies and neurophysiological results.
Conventional robotic control methods solve predetermined kinematic equations to produce the desired motion of the end effector. In many systems, this is not possible due to unknown or changing dynamics. Human infants have little self-knowledge at birth and must learn to control their own limbs almost entirely from scratch. In this experiment, a motor babbling process is used to bootstrap a model of hand-eye coordination on the iCub humanoid robot. The robot’s visual space is divided into receptive fields, each of which learns a SOM of the corresponding arm joint angles. The visuomotor learning model is evaluated using two different hand tracking methods: noisy truth to approximate an optimal tracking method, and a simple developmentally plausible tracking method. In both cases, the model successfully reduces the reaching error after learning. In the case of the model trained with the noisy truth, median reach error across visual units converges to 4.3 cm, from a median error of 22.3 cm in the untrained model. The simple visual tracking method produced a median error of 17.8 cm, with potential to converge to a lower error given long training times.

3.1 Training Procedure

3.1.1 Motor Babbling

The basis of motor babbling on the iCub is the generation of random joint configurations. Although the arm has 16 degrees of freedom, to keep complexity of the model as low as possible only four are used: three for the shoulder and one for the elbow. In theory, the model presented here could be used to learn a mapping from the joints of the wrist and fingers to hand pose. Such a model could be learned entirely independently of the model
Joint configurations generated during babbling are tested for safety using the exact kinematics of the arm before being executed. Due to the delicate construction of the robot, it is essential to ensure that self-collisions do not occur. For this reason, while the left arm babbles, the right arm is extended away from the body. The Cartesian location of the generated joint configuration is tested to ensure that the location in the $x-y$ plane is at least 30 cm from the origin of the robot coordinate frame, shown in Figure 3.1.

Figure 3.1: Cartesian reference frame for iCub, from [101]

Two methods were tested for motor babbling. The first was a random walk across the joint space of the arm, starting from a random configuration. This method did not produce an even distribution of hand positions across reachable space and left regions near the boundary poorly trained. The method used in the results presented here is a Monte Carlo style generation of joint configurations. At each step, one joint is increased by a random amount while the remaining are fixed, producing a sampling of reachable space after several cycles. Figure 3.2 shows the region which is reachable by the robot, obtained from the visual receptive units which had been visited by the hand after 50,000 training samples.

3.1.2 Hand Localization

Visuomotor models were trained with two methods of visual tracking. First, to evaluate the visuomotor model independently of hand localization, a noisy version of the true hand position was tracked. At each time step, the robot fixated on the hand location with noise drawn from a $5 \text{ cm}^3$ uniform distribution added. This size is roughly equivalent to the dimensions of the iCub’s
hand. Even with an ideal hand detector, it is unlikely that a consistent part of the hand would be located every time. For this reason, it was deemed appropriate to test the model by training on noisy truth, to ensure that the model is robust enough to learn despite inevitable noise in hand tracking.

The second method used to detect the hand was motivated by the observation that infants begin to track their own hand visually before they understand that they own the hand. To evaluate the performance of the model under the circumstances of incomplete self-knowledge and rudimentary attentional systems, a simple visual hand localization method was developed. In this method, the robot visually tracks anything which satisfies three requirements: possessing edges and corners, moving, and close to the face. This behavior can be seen in infants, regardless of the object of interest being tracked. In this experiment, such requirements result in the hand being tracked preferentially. The Shi-Tomasi corner detection algorithm is used to generate the location of corners [102], which is then used as features to calculate the pyramidal Lucas-Kanade optical flow [37]. Regions of high optical flow which are within the bounding box of reachable space are assigned
a high salience. Using the salience for each eye, the iCub gaze controller fixates on the estimated hand location [103].

3.2 Visuomotor Mapping

A model of viusomotor learning was chosen according to several criteria. First, it must implement a developmentally plausible learning method. Second, it must be able to learn the joint to end effector mapping as an under-determined system. Finally, it must be readily usable for control.

3.2.1 Motor SOM for Kinematic Modeling

The visual space is populated by units $V_{ijk}$, each with a receptive field size of $5 \text{ cm}^3$. Since the iCub gaze controller is inherently Cartesian, the visual space is also defined as such. In a truly developmental system, the mapping between both gaze and motor configuration to an invariant egocentric space would be learned. At each time step, the visual unit which corresponds to the detected hand location is trained. Due to the surjective nature of the mapping from joint angles to visual location, facilitating smooth motion requires that multiple joint configurations be learned for each unit. To address this problem, a single-dimensional SOM is used to learn a manifold of joint configurations corresponding to the visual unit. In this experiment, the SOM is four motor units arranged in a circle, as seen in Figure 3.3.

![Visuomotor model](image)

Figure 3.3: Visuomotor model in which each visual receptive field maps to a 1D motor SOM
Each motor unit is initialized with a random joint configuration. Once a visual unit has been activated by hand detection, the winning motor unit is chosen according to the standard rule:

\[ y_n = w_n^T x \]  

where \( y_n \) is the output of SOM neuron \( n \), \( w_n \) is its weight, and \( x \) is the vector of joint angles. The update rule for the weights is

\[ \Delta w_n = \eta(x - w_n)\Lambda(n, n^*) \]  

where \( \eta \) is the learning rate, \( n^* \) is the unit with the highest activation, and \( \Lambda(\cdot, \cdot) \) is a neighborhood function. Here, \( \eta \) is chosen to decay with the training count according to:

\[ \eta(count) = 0.5e^{-\text{count}/(10 \times \text{totalMotorUnits})} \]  

The neighborhood function is chosen as:

\[ \Lambda(n, n^*) = \begin{cases} 
1, & |n - n^*| = 0 \\
0.25, & |n - n^*| = 1 \\
0, & \text{else}
\end{cases} \]  

Algorithm 3.1 shows the training procedure.

**Algorithm 3.1 Visuomotor Model Acquisition**

```
loop
  move arm randomly
  if hand visible then
    fixate on hand
    select visual unit
    for all units in corresponding motor SOM do
      calculate unit activation
    end for
    select winning motor unit with highest activation
    update weight of winning unit and neighbors
  end if
end loop
```
3.2.2 Model Performance

The noisy truth model was run for 50,000 samples, and the optical flow hand localization model for 25,000 samples. Since acquiring data for the optical flow model is more time consuming, due to the possibility of failure to acquire fixation on a salient point, fewer samples were used. For comparison, an untrained model consisting of the randomly initialized joint angles was saved, here referred to as the null model. Due to the quantizing nature of this spatial model, there is an expected sampling error. For the given resolution, 5 cm³ per unit, it is 2.4 cm. This is the expected difference between the center of the visual receptive field and the hand location corresponding to the motor unit’s tuning, given perfect hand localization during training. Due to the noise in any realistic hand localization technique, errors this low are not to be expected in the trained model. The following analyses all concern the average error across all four motor units corresponding to given visual unit. Figure 3.4 shows a visualization of unit errors for a slice through the visual space at \( z = 0.1 \) m.

It can be seen that high errors persist near the edges of reachable space, possibly due to these units not being truly reachable, but rather trained due to error in hand localization. In the bulk of the reachable space, however, the error drops to very low levels with sufficient training. Figure 3.5 shows the median error over time for both models.

In Figure 3.6, histograms of the errors for all reachable visual units are shown. As expected, the null model displays an approximately normal distribution of errors, while the noisy truth model displays a strong leftward skew toward low errors. Figure 3.7 shows box plots of the null model and the noisy truth model. As can be seen, these distributions are different, indicating that the model is correctly learning visuomotor correspondences. Figure 3.8 compares the null model to the optical flow model after 25,000 samples. In this case, there is a difference in the distributions, but less stark.

Due to the topology of the reachable space, it is assumed that the prior distribution of errors varies greatly across visual units, with more central units having lower expected errors for the untrained model than ones in the perimeter. For this reason, a paired difference test comparing the null model to the trained model was chosen as an appropriate measure of statistical significance of the changes due to the learning process. The Wilcoxon signed
Figure 3.4: Slices of errors on $x - y$ plane at $z = 0.1$ m, grey areas indicate region is not in reachable space
Figure 3.5: Median error vs. number of samples, dashed line shows lower bound on error for receptive field size
Figure 3.6: Histograms of errors for visual units
Figure 3.7: Comparison of null model to noisy truth model after 50 k samples, dot is mean, line is median

Figure 3.8: Comparison of null model to optical flow model after 25 k samples, dot is mean, line is median
rank test was chosen to avoid assumptions about the types of distributions of the errors [104]. For the noisy truth model after 50,000 samples, \( p = 1 \times 10^{-48} \). For the optical flow model after 25,000 samples, \( p = 6 \times 10^{-15} \). In both cases, the p-values indicate a meaningful reduction in error from training.

### 3.3 Dynamic Reaching

The visuomotor mapping can be used to implement a velocity controller for the robot. This method has been previously demonstrated on a simpler robotic arm in a 2D task space [77]. Define the set of all SOM weights from the whole visual array as \((m_0, m_1, \ldots, m_K)\), the proprioceptive attractors. The function used for control, in terms of these attractors and the current joint angles \( x \) is

\[
f(x) = \sum_{k=0}^{K} V(m_k) N(|x - m_k|)(m_k - x)
\]  \hspace{1cm} (3.5)

\( V(m_k) \) is the activation of the visual neuron to which \( m_k \) corresponds. Activation is determined by salience, such as object or hand detection. \( V(m_k) \in [0, 1] \), where activation can be either binary or continuous in the interval. Discussion on the implications of these two options is presented in Chapter 4. \( N(|x - m_k|) \) is defined as

\[
N(|x - m_k|) = \frac{G(|x - m_k|^2)}{\sum_{k'} V(m_{k'})G(|x - m_{k'}|^2)}
\]  \hspace{1cm} (3.6)

\( G(x) = e^{-\beta x^2} \), where \( \beta \) is chosen to control the slope of the attraction basins. \( f(x) \) can now be used to define a velocity control law:

\[
\tau \ddot{x} = A f(x) + \epsilon
\]  \hspace{1cm} (3.7)

The behavior of the arm can be altered by adjusting \( A \) and \( \epsilon \), the gain and noise. High gain and low noise cause the arm to reach directly for the nearest active visual unit. Low gain and high noise trigger random exploration. In the case of defining a single salient point, as in many reaching applications,
the control law simplifies to:

$$\tau \dot{x} = A(m_k - x) + \epsilon$$  \hspace{1cm} (3.8)

In this case, the system behaves as a simple point attractor. Algorithm 3.2 shows the learned reaching procedure.

<table>
<thead>
<tr>
<th>Algorithm 3.2 Learned Reaching</th>
</tr>
</thead>
<tbody>
<tr>
<td>loop</td>
</tr>
<tr>
<td>get visual fixation</td>
</tr>
<tr>
<td>if fixation point unreachable then</td>
</tr>
<tr>
<td>calculated nearest reachable point</td>
</tr>
<tr>
<td>end if</td>
</tr>
<tr>
<td>get current arm joint values</td>
</tr>
<tr>
<td>get active visual unit</td>
</tr>
<tr>
<td>for all motor units in corresponding SOM do</td>
</tr>
<tr>
<td>calculate motor unit activation</td>
</tr>
<tr>
<td>end for</td>
</tr>
<tr>
<td>get weights of winning motor unit</td>
</tr>
<tr>
<td>set arm joint velocities according to Equation 3.7</td>
</tr>
<tr>
<td>end loop</td>
</tr>
</tbody>
</table>

3.4 Discussion

The resulting accuracy of the learned visuomotor map indicates that this model is a sound method for acquiring hand-eye coordination. The performance of the model is strongly dependent on the method for hand tracking, as can be seen in the comparison between the two versions of localization. However, even with the extremely simple and error-prone optical flow method, training results in statistically significant learning. The maximum number of samples used here were collected over a fraction of the time which is used by human infants during a comparable developmental process. It is likely that the high noise in the optical flow model slows convergence but does not fundamentally counteract the effectiveness of the learning method, and that this model would also converge to a lower error if allowed to run on developmental time scales.

In a learned egocentric visual model, lateral activation among the visual units is expected. Given a sufficiently accurate hand localization model, such
lateral activation could serve to further enforce the topological organization of the visuomotor space. It is possible that in a developmental context, the mapping between egocentric visual fixation and head motor activity would be learned before reach acquisition. The learned model only encodes an open loop reach. In adult human reaching, both open and closed loop control occurs, the latter being used to correct error from the former. For greater reach accuracy, the model presented here could be combined with visual feedback, more closely mimicking actual human motor behavior. This learning method can be used continuously, allowing the robot to adapt to any potential changes in dynamics during operation, and even potentially be applied to learning the dynamics of tool use.
CHAPTER 4

MOTOR PRIMITIVES FOR ASSOCIATIVE LEARNING

Representation of an egocentric space is a prerequisite for spatial reasoning. The outlined visuomotor model provides a mechanism for learning of simple coordination and reaching. By providing an interface between motor control and a spatial representation, it can also serve as a basis for semantic motor learning. Human motor behavior is generated by the combination of motor primitives, which are velocity fields in the task space [62]. The model presented in Chapter 3 provides a biologically plausible method for learning these primitives. In this chapter, the features of the model are discussed in the context of full body motion and cross-modal learning. Additionally, a method for learning grasp types corresponding to objects using the motor model is presented.

4.1 Features of Self-Organizing Motor Model

The self-organizing motor model has many features which make it suitable for cognitive developmental robotics. It is unsupervised, the robot can continue training while behaving, and requires no a priori information. The accuracy of the model is essentially only limited by the size of the visuomotor receptive fields, as demonstrated in Section 3.2. In addition to the model’s advantages for hand-eye coordination acquisition, it has features which allow it to be incorporated into a broader developmental learning system.

4.1.1 Implicit Continuity of Representation

The units in the model described in Section 3.2 were given non-overlapping receptive fields to allow the motor SOM to be evaluated in isolation. In the parietal cortex, where motor and retinal reference frames are integrated for
egocentric motion planning, spatial receptive fields overlap [105]. A more biologically realistic implementation of the visuomotor model would train motor SOM corresponding to the neighbors of the winning unit. In the human brain, there is no reason to believe that any topological representation is present at the beginning of development. In the case of the robot, it is possible to learn the spatial tuning of the visual units themselves, creating a hierarchical SOM. Training problems associated with multilayer SOM can be addressed through developmental staging, training components of the model individually from the sensory periphery “inward”. Topologically organized units with Gaussian receptive fields produce equivalent activation levels to a SOM with a neighborhood function, a subject explored in depth in Chapter 5. The result of a model where visual units have overlapping receptive fields is an implicitly continuous representation of egocentric space. This is an advantage when incorporating the motor representation into an associative memory framework.

4.1.2 Model Decoupling

Complex motor behavior in humans is generated by superimposing primitives represented as vector fields [60]. The dynamics of the human body allow decoupled superimposable kinematic models of various muscle systems. One notable kinematically decoupled system is human hand orientation and location, an actuator design commonly reproduced in robotics. In addition, the torso pose and finger positions are decoupled from the arm. The self-organizing motor model can be used to learn efficient representations of whole body movement by exploiting this decoupling. Like in humans, separate self-organizing models can be learned for each of the dynamic subsystems of the robot and superimposed. In the experiment from Section 3.2, only hand position was learned. However, a complete model would need to learn wrist and finger pose to be used in object manipulation tasks.

4.2 Grasp Acquisition

Grasp acquisition is bootstrapped by the grasp reflex present in infants. When object manipulation begins, grasp types are associated with object
properties to form a model of effective grasping strategies. Especially in the early stages of infant grasp acquisition, haptic feedback is central to evaluating the effectiveness and stability of grasp attempts [106]. The iCub robot in the Language Acquisition and Robotics Group does not have haptic sensors, making developmental grasping experiments difficult. The ability of the robot to detect its grip on the object is important for autonomous object exploration. However, outlined here is a method for grasp acquisition based on established methods which could be implemented after the addition of haptic sensors.

The model from Section 3.2 can be used to learn and represent primitives of the location of the fingertips relative to the palm of the hand. If the locations of the fingers are treated as implicitly continuous, they can be provided as observed states to a continuous observation HMM. The unsupervised clustering properties of the HMM allow different grasp types to be encoded as hidden states. Within a given category of grasp, the primary variation in finger position is due to differing object sizes. The HMM is able to account for object size differences in the variance of Gaussian observation distributions. It is expected that the largest variance in the distribution for a given finger would be along an axis connecting the center of the object to the finger, illustrated in Figure 4.1.

Cascaded Hidden Markov models are generally effective for the implementation of multimodal associative memory. In the Language Acquisition and Robotics Group, it has been used to learn both associative models of visual objects and speech and models of hand trajectory and speech [6, 7]. A similar framework is proposed here to learn object-grasp associations. The motor observation stream consists of samples of continuous valued fingertip positions, \( A_n = [f_0, f_1, f_2, f_3, f_4] \), learned by a visuomotor acquisition model. The continuous-valued visual observations \( B_n \) consist of eccentricity and first moments of objects segmented using the method from [6]. The hand pose HMM, \( M_A \), learns a set of grasp types, \( a_i \). The visual HMM, \( M_B \), learns shape categories, \( b_i \). The concept HMM, \( M_C \), takes symbolic observations in the form of the states \( a_i \) and \( b_i \), learning hidden states \( c_i \). The states \( c_i \) are not an observable process in the system, but they encode grasp/object pairs. Figure 4.2 shows the model structure.

The system is trained using RMLE to allow for online learning, as described in [6]. During training, the robot reaches for the detected object and
executes the grasp method predicted by the current model given the visual classification. The top-level HMM is only trained while the object is in the hand of the robot, as detected by haptic sensors. This approach allows temporarily successful but unstable grasps to be used for training, while still putting the most training emphasis on the most stable grasps. Early in the training process, it is unlikely that any grasp will be stable; this approach allows unstable but successful grasps to be used as first approximations. An outline of the training method is shown in algorithm 4.1.

**Algorithm 4.1 Grasp Acquisition**

```plaintext
while object visible do
    if object in contact with haptic sensors then
        use RMLE to train models $\mathcal{M}_A$, $\mathcal{M}_B$, $\mathcal{M}_C$
    else
        reach for object
        classify object using $\mathcal{M}_B$
        generate grasp type $a_i$ given current model $\mathcal{M}_C$
        generate sample finger positions for $a_i$, attempt grasp
    end if
end while
```

Figure 4.1: Variance in finger position for a given grasp style
Figure 4.2: Grasp acquisition architecture
CHAPTER 5

TOPOLOGICAL MAPPINGS FROM
NEURONAL DYNAMICS

The self-organizing map (SOM) is often regarded as a computational model of topological learning in the brain. However, the neighborhood function, essential to its behavior, has not been explained in terms of observed neural behavior [107]. In this chapter, a biologically based model of the SOM is presented.

5.1 Scale-Free Networks of Neurons

Groups of neurons are thought to form scale-free networks, characterized by connectivity distributed according to

\[ p(k) \sim k^{-\gamma} \] (5.1)

where \( p(k) \) is the probability of a node having \( k \) connections. Typically, \( 2 < \gamma < 3 \). In the case of the Barabsi-Albert generative model for scale-free networks, \( \gamma = 3 \) [108]. The expected connection distance between two units in such a network consisting of \( N \) units is

\[ \ell \sim \frac{\text{ln}(N)}{\text{ln}(\text{ln}(N))} \] (5.2)

For example, in a network of 1000 neurons connected according to this model, the expected number of synapses required for a signal to reach one neuron to another is fewer than four. This model of local connectivity is appealing due to its consequences for propagation of activity. Simulations of scale-free networks of Hodgkin-Huxley neurons have demonstrated that such populations synchronize their firing phase under stimulation [50]. Borrowing terminology from that study, these synchronizing populations will be referred to as “components”. It was further demonstrated that sparsely interconnected groups
of components have the potential to encode memory states in their dynamics. It is commonly understood that in most neurally inspired computational methods, such as SOM and backpropagation networks, units are not analogous to individual neurons but rather to populations. Due to their many appealing properties, the neuron components described here are a suitable analogy to the canonical SOM unit.

5.2 Plasticity Models

Synaptic plasticity is central to learning and memory. Hebbian models, while suited to certain computational tasks such as component analysis, assume exact simultaneity of presynaptic and postsynaptic activity. Spike-timing dependent plasticity models allow varying levels of potentiation and depression of synaptic strength according to temporal proximity of presynaptic and postsynaptic activity. The functional form of STDP is varied, both between types of neurons and within the modes of an individual neuron. One basic computational model of STDP has been used to implement competitive Hebbian-like learning [109]. Here, the same simple model will be used:

$$\Delta w_n = \begin{cases} w_p e^{\frac{-\Delta t}{\tau_p}}, & \Delta t > 0 \\ -w_d e^{\frac{-\Delta t}{\tau_d}}, & \Delta t \leq 0 \end{cases}$$

(5.3)

The convention used here is that $\Delta t$ is positive when the presynaptic spike precedes the postsynaptic spike. The choices of $w_p$, $w_d$, $\tau_p$, and $\tau_d$ effect the network dynamics.

5.3 Topology-Learning Neural Network

Based on the introduced principles, the SOM can be interpreted as a network of neural components subject to STDP. Let $N_{pre}$ be a group of components with differing receptive fields. These receptive fields might correspond to retinal locations, motor configurations, or any other space suited to topological representation. Let $N_{post}$ be another group of components with lateral sparse connections to their neighbors. Initially all components of $N_{pre}$ synapse sparsely onto all components of $N_{post}$ with low random synapse
strengths. It is assumed that intergroup component distance is significantly larger than intragroup component distance, as would be the case when groups are in different brain regions. Furthermore, it is assumed that distance is the most significant factor in determining spike transmission speed. Figure 5.1 shows the proposed network.

Figure 5.1: Connection between two groups of components, red shows lateral connections

Since all neurons in a component have the same base firing rate, synchrony is chosen as a proxy for “activation” of a unit from the SOM. Due to the properties of scale-free networks, it can be assumed that the neurons within a component which form intercomponent connections are likely to be “hubs” in the component. Therefore, it is appropriate to expect that for significantly strong synapses between components, synchronization in the input component will trigger the same in the output component. In [50], the state, $\theta$, was defined as the variance of the phase of firing neurons in the component. Here, a synchrony variable is defined:

$$\zeta = 1 - \frac{\theta}{2\pi}$$

(5.4)

Using this metric, synchrony is always nonzero. The minimum synchrony, when phase is distributed uniformly, is $\zeta_{\text{min}} = 0.26$. In a real neural network, some amount of synchrony is inevitable, even in the absence of external stimulation. Potentiation events are assumed to correspond to the mean phase of the input component, so the probability of an individual input neuron having fired is proportional to the synchronization:

$$p(n \in N^{i}_{\text{pre}}) \propto \zeta_{\text{pre}}^i$$

(5.5)
Likewise, the probability of an individual output neuron activating due to this event is:

\[
p(m \in \mathcal{N}^j_{\text{post}} | n \in \mathcal{N}^i_{\text{pre}}) \propto w_{nm}
\]  

(5.6)

From this, we derive an estimate of the probability that an individual synapse will be potentiated:

\[
p(w_{nm} \uparrow) \propto w_{nm} \zeta^i_{\text{pre}}
\]  

(5.7)

The synchronization of \( \mathcal{N}^j_{\text{post}} \) is again proportional to the probability of a postsynaptic neuron firing during the potentiation event, so it can be defined as

\[
\zeta^j_{\text{post}} = \bar{w}_{ij} \zeta^i_{\text{pre}}
\]  

(5.8)

where \( \bar{w}_{ij} \) is the average synaptic weight between the two components. In [50], relationships between components are analyzed as ODEs. However, here we discuss the steady-state synchrony levels of two connected components.

Let \( \Delta t_f \) be the intergroup synapse time and \( \Delta t_n \) be the intragroup synapse time. From Equation 5.3, the change in strength of a potentiated synapse \( nm \) is

\[
\Delta w_{nm} = w_p e^{-\frac{\Delta t_f}{\tau_p}}
\]  

(5.9)

From this, the average synaptic strength change between \( \mathcal{N}^i_{\text{pre}} \) and \( \mathcal{N}^j_{\text{post}} \) is

\[
\Delta \bar{w}_{ij} = \bar{w}_{ij} \zeta^i_{\text{pre}} w_p e^{-\frac{\Delta t_f}{\tau_p}}
\]  

(5.10)

The immediate neighbors in the output group, \( \mathcal{N}^k_{\text{post}} \), are activated by lateral connectivity, possibly in addition to weak influence from the input group. As before, the probability that an individual synapse \( np \) from \( \mathcal{N}^i_{\text{pre}} \) to \( \mathcal{N}^k_{\text{post}} \) is activated and thus potentiated is:

\[
p(w_{np} \uparrow) \propto \bar{w}_{\text{lat}} \zeta^j_{\text{post}}
\]  

(5.11)

where \( \bar{w}_{\text{lat}} \) is the average intragroup synaptic weight between output components. If the neurons in this group do not fire until stimulated by lateral activity, the change in average synaptic strengths from the input component to these components is

\[
\Delta \bar{w}_{ik} = \bar{w}_{\text{lat}} \zeta^j_{\text{post}} w_p e^{-\frac{\Delta t_f + \Delta t_n}{\tau_p}}
\]  

(5.12)
Propagation might continue a few groups further, but will eventually stop due to attenuation across lateral connections. In general, for a component \( N^k_{post} \) which is \( \ell \) components away from \( N^j_{post} \) which was initially activated, the update of average synaptic strength from \( N^i_{pre} \) is

\[
\Delta \bar{w}_{ik} = \Delta \bar{w}_{ij} w_{lat} e^{-\frac{\Delta t}{\tau_p}} \tag{5.13}
\]

This rule is a form of the neighborhood function used in the SOM. With \( \frac{\Delta t}{\tau_p} \) chosen to be 0.05, Figure 5.2 shows the effect of choice of \( w_{lat} \) on the neighborhood function. Assuming the previously discussed \( \zeta_{min} \), the probability of a synapse being spuriously activated right before input synchronization, and thus depressed, is

\[
p(w_{np} \downarrow) = \zeta_{min}\zeta_{pre} \tag{5.14}
\]

From this, it can be inferred that the expected depressive effect at each activation of an input component is

\[
\Delta \bar{w}^d_{ij} = -w_d \zeta_{min}\zeta_{pre} e^{-\frac{\Delta t_f}{\tau_d}} \tag{5.15}
\]

Given appropriate parameters of the model, potentiation will predominate.
for highly synchronous postsynaptic components and their neighbors, while other components will be weakly depressed. This has the effect of “pruning” connections and further focusing the topological organization of the output group.

5.4 Discussion

The model presented here is a hypothesis about how an SOM might be implemented in a real neural circuit. With the group of input components standing in for a real-valued vector as used in the typical SOM, it fulfills the desired properties of adjacent output units representing similar inputs. A neighborhood update rule causes adjacent output components to receive strong synapses from similar input components. Inherent in the model described here is a simplification of the synchrony of a component. Here, individual neuron spikes are treated as entirely synchronous with the mean or too asynchronous to contribute to potentiation. In a real neural circuit, the specific timing of spikes, even near the mean, affects potentiation. Since the model presented here deals with averages over large numbers of neurons, such effects are discounted. Although this model successfully derives the SOM from biological phenomena, simulation is needed to confirm its representational power. The parameters presented here depend on experimental values, it is possible that they might differ depending on the type of neuron and chemical signaling. The form of STDP itself varies widely, it is likely that different forms lead to distinct computational modes of networks. Further study of the emergence of computational processes from spiking neural networks is certain to contribute to understanding of the biological basis of cognition.
CHAPTER 6

CONCLUSION

While much work is being done on both neural network modeling and behavioral modeling, connections drawn between the two are still limited. Cognitive developmental robotics provides a method for evaluating the results of proposed behavioral models, while neural dynamics suggest possible implementations of observed processes. This work has attempted to unite these two perspectives on cognition.

6.1 Mapping between Sensory Reference Frames

The hand-eye coordination model presented in Chapter 3 demonstrates the capability of a biologically plausible algorithm to learn a visuomotor mapping. Results for an ideal hand localization method were near the minimum theoretical error for the model, on the order of the hand size itself. Even for a rudimentary hand localization model based on inborn attentional preferences, the trained model was significantly superior to the initial state of the system. This model was used for a control system which generates smooth trajectories as the hand performs a reaching action. Generally, this method allows different sensory reference frames to be mapped onto one another. In this implementation, the joint values were mapped onto a Cartesian representation of the egocentric visual space. In an entirely biologically plausible system, this would not be possible. It is unclear whether motor activity would be mapped onto a representation of polar visual fixation or if both visual fixation and motor activity would map onto a third latent topological representation, possibly similar to Cartesian. Due to kinematic decoupling of humanoid bodies, it is possible to create complex motor behavior via the superposition of motor primitives learned for several dynamical subsystems. Although it was outside the scope of these experiments, similar motor models
could be developed for the torso, wrist orientation, and finger pose.

6.2 SOM as Model of Topological Learning

The SOM has become an established data visualization and pattern classification technique. It was initially conceived as a model of the formation of topological maps in the brain. However, it has not yet been established how the neighborhood function, essential to topological organization in the SOM, might be implemented in a realistic neural network. In Chapter 5, a network of Hodgkin-Huxley neurons is proposed which learns a topological mapping of the input. This model shares central properties of the SOM: competitive learning and topological mapping of the input space. Many machine learning techniques exist which mimic high-level human learning, but few have been explained in terms of neural models. Continued interaction between the computational neuroscience and machine learning communities will hopefully lead to an expansion of this line of inquiry. Much as results of computational neuroscience have informed the models developed here, machine learning techniques which replicate observed human development can serve as starting points for models of neural circuits.

6.3 Associative Memory

Previous studies have been done demonstrating the ability of established associative memory models to learn motor representation [8, 7]. These experiments have been based on a priori knowledge of robot kinematics. The model from Chapter 3 has many properties suitable to use in a cascaded associative memory. Presented here is a foundation for integration of a fully learned primitive-based motor model with a multimodal CHMM associative memory. The example of grasp acquisition presented in Chapter 4 is a simple application of the motor model in the context of multimodal learning. The ability to learn concepts or symbols which span motor activity and other modalities has many applications. Acquisition of motor affordances can potentially be modeled using this framework. In addition, the motor theory of language predicts that phonemes are understood as an observable process of
underlying position of the vocal apparatus. This model would be a promising approach to learning the association between motor activity of the vocal tract and generated sounds.

6.4 Future Work

There are many improvements that can be made to the hand-eye coordination model. The rudimentary hand localization model would ideally be replaced by a developmentally plausible but reasonably accurate one. Given the good results for an ideal method, it should be possible to develop a system which accurately acquires hand-eye coordination without a priori information. It is reasonable to assume that all cortical topological mappings are learned, with the only hardwired mappings existing where the topology has an inherent relation to the physical world, such as the retina. In a more complete model of visuomotor coordination, the mapping from retinotopic coordinates and head and eye pose onto an egocentric spatial representation would also be learned. In such a model, training of the motor SOM can be done with lateral visual activation, ensuring that contiguous visual units have similar motor tuning.

There are many applications to semantic associative memory which could be implemented combining these results and previous memory models. Grasp acquisition according to object shape is an immediate application. A study on the motor theory of language could be done using this framework with a robotic version of the human vocal tract. The description of neural formation of topological maps in Chapter 5 mirrors the SOM, but simulations are required to evaluate its behavior in comparison to similar machine learning models.

6.5 Platform-Independent Model of Intelligence

Because all scientists are human, human cognition is of intense interest in the scientific community. However, none of the results discussed here are unique to humans. The author’s dog reaches for visual targets, does simple manipulation of objects, and forms topological cortical maps. Even very
simple organisms with nervous systems display behavior which is challenging at best to implement with modern artificial intelligence and robotics technology. Models of behaviors unique to humans, such as language, are part of the story, but we still lack adequate models of how basic cognitive processing emerges from the vertebrate nervous system.

In essence, an intelligent system is one which predicts the outcome of its own state and actions on the observed environment and acts to maximize a goal. In living systems, the goal is to stay alive long enough to reproduce. However, this paradigm can be applied to machine learning systems, such as learning to control a prosthetic limb accurately. It is difficult to quantify what makes a system truly intelligent. Many modern technologies achieve excellent classification and recognition results in image and speech processing due to the high availability of training data. However, few consider these statistical methods to be intelligent in the sense of biological cognition. Two of the defining factors in the common interpretation of intelligence are robustness and computational efficiency. In machine learning, these two factors are often in opposition. In biological cognition, both are fulfilled. When evaluating by both of these standards, no modern robotic system can compare to even the simplest vertebrate. It is worth noting that the simplest forms of life display highly stereotyped stimulus response; this has a parallel to less robust machine learning techniques. Hard-coded versus learned processing exist on a continuum in intelligent systems. Many evolutionary factors weigh into the balance of these two modes; one significant trend is a greater reliance on learning with increased organism lifespan.

Future research into cognitive robotics might do well to study the dazzling variety of intelligent systems observed on this planet. From frogs to termite colonies, there are many seemingly simple organisms which have much to teach us about intelligent processes. The class of artificial systems which can be described as intelligent will continue to grow, and the need continues to develop a platform-independent model of such systems.
REFERENCES


[26] J. Miller and P. Lommel, “Biomimetic sensory abstraction using hier-


[28] R. Rizzo and A. Chella, “A comparison between habituation and con-
cience mechanism in self-organizing maps,” IEEE Trans. Neural Net-

[29] M. V. Hulle, “Kernel-based equiprobabilistic topographic map forma-


[31] L. Baum, “An inequality and associated maximization technique in
statistical estimation for probabilistic functions of a Markov process,”


[35] J. Kamarainen, V. Kyrki, and H. Kalviainen, “Invariance properties of


[37] J. yves Bouguet, “Pyramidal implementation of the Lucas Kanade fea-

[38] R. Kalman, “A new approach to linear filtering and prediction prob-


