Sensei: A Real-Time Recognition, Feedback and Training System for T'ai Chi Gestures

by

David Alan Becker

B.A., Astrophysics and Astronomy and Physics
Harvard University, Cambridge, MA
June 1993

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of
MASTER OF SCIENCE IN MEDIA TECHNOLOGY
at the
Massachusetts Institute of Technology
May 1997

Signature of Author ______________________________
Program in Media Arts and Sciences
May 9, 1997

Certified by ______________________________
Alex Pentland
Toshiba Professor of Media Arts and Sciences
Academic Head, Media Laboratory
Thesis Supervisor

Accepted by ______________________________
Stephen A. Benton
Chairperson
Departmental Committee on Graduate Students
Program in Media Arts and Sciences

Sensei: A Real-Time Recognition, Feedback and Training System for T'ai Chi Gestures

by
David Alan Becker

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning
on May 9, 1997
in partial fulfillment of the requirements for the degree of

Master of Science in Media Technology

Abstract

The real-time, interactive feedback system we developed, *Sensei: The T'ai Chi Teacher*, is presented. This system provides a good platform on which to build a more sophisticated teaching, training, and feedback tool for gestures or action. In this document, we hope to substantiate that thesis by showing that *Sensei* does have the components necessary to be a good foundation. First of all, it is capable of performing real-time, multiple user, gesture recognition. Users of the system are free to practice *T'ai Chi* gestures, starting anywhere in the sequence of moves, and the system recognizes their actions. In addition, a complete teaching system must be able to give both positive and critical feedback to the user. This ability implies a knowledge of the instants in the user's performance of a gesture where the user was both least and most accurate in the movement. Both tasks are accomplished through the use of Hidden Markov Models. Experiments testing these abilities are presented. The work concludes with a discussion of future extensions.

Thesis Supervisor: Alex Pentland
Title: Toshiba Professor of Media Arts and Sciences

Sensei: A Real-Time Recognition, Feedback and Training System for T'ai Chi Gestures

by
David Alan Becker

The following people served as readers for this thesis:

Reader: ____

Rosalind Picard
NEC Development Professor of Computers and Communications
Associate Professor of Media Technology
Media Laboratory, Massachusetts Institute of Technology

Reader: ____

Billy L. Little
Founder
Cancer Support Center of St. Louis

Acknowledgments

More people than I have room to acknowledge here have played roles in helping me to be both intellectually and physically capable of producing this thesis. From those whom I neglect to mention, I humbly ask for forgiveness and for understanding; your help is appreciated, even if I have not directly stated so.

Among those, however, whom I cannot fail to explicitly thank is my advisor, Sandy Pentland. Sandy has provided me with everything one might expect from an advisor: guidance, support, and vision. More importantly to me, however, he has been extremely gracious and understanding as I've recovered from illness, and for that, as well, I thank him.

My first reader, Bill Little, has taught me much about living a healthy life, a skill for which I can hardly repay him. Roz Picard, my other reader, not only gave me excellent feedback and comments, but has been both a creative and caring influence during my sojourn at the Media Lab.

Many, many of my fellow students at the lab deserve my sincere gratitude. To Chris Wren I am deeply indebted for all the answers he has provided to the myriad questions I incessantly ask him. It would have taken Chris less time to simply implement all of my code on his own than it did to field all of my queries. However, this would have freed me to develop my Diablo character more than his, and how could he allow that?

I am fortunate to have shared an office with Andy Wilson for the last three years. Not only does he tolerate my daily interruptions of his work, but he foresaw every technical difficulty I would have along the way. As such, he generally had an answer for any question I would throw at him - usually before the words had escaped my mouth.

Lee Campbell and Ali Azarbayejani have been the best of colleagues – willing and capable of general help and great work partners. Chester Barberra, a veritable icon among Vismodders, has done a superb job of keeping everything in our group in perspective. Plus, he provided us all with all the free bags of Cheetos we could eat. Lastly, to Sumit I owe thanks not only for answering questions, but for being the best piano teacher in the world.

The rest of the entire Vismod gang has made the last three years exciting and stimulating. They are a group I will not readily forget.

My mother and father have been wonderfully supportive over the years. Their unwa-

vering love is something for which I can never thank them enough.

Finally, there is also the woman that I can't believe I have the good fortune of being married to, Stephani. Steph, you have given me everything I could ever want in my life; writing a thesis is such a small thing compared to that. So to you, I simply say thanks for being you.

Contents

1 Introduction **10**

1.1 Statement of Thesis 10

1.2 Motivation 10

1.2.1 Staying Alive 11

1.2.2 Sensei: the T'ai Chi Teacher 12

1.3 Approach 12

1.4 Other Applications 14

1.5 Outline of Thesis 14

2 Related Work **15**

2.1 Imagery and Relaxation 15

2.1.1 Early Work in Imagery, Visualization, and Hypnosis 15

2.1.2 Imagery as Applied to Cancer 16

2.2 Using HMM's for Gesture Recognition 18

3 Implementation of the Recognition System **20**

3.1 Description of Hidden Markov Models 20

3.2 Implementation of HMM's 21

3.2.1 Topology 21

3.2.2 Choosing the observation vectors 22

3.2.3 Training the model 22

3.2.4 Performing the recognition 23

3.3 Implementing the T'ai Chi Teacher's HMM's 23

3.3.1 Observation vectors 23

3.3.2 Choosing a Topology for the T'ai Chi Teacher 26

3.3.3 Training the Models 27

4 Recognition of Gestures and Feedback 30

4.1 Gesture Recognition 30

4.2 Localizing Salient Moments in the Gesture 32

5 Experiments 34

5.1 Gesture Recognition Experiment 34

5.2 Identifying Salient Points in the Gesture 35

5.3 Analysis of Results 38

6 Summary and Future Work 42

6.1 Summary 42

6.2 Future Work 42

A Computational Advantage of Training the Models Backwards 44

B Smoothing the Data 45

B.1 Filtering 45

B.2 Resampling 45

List of Figures

1-1 The five *T'ai Chi* gestures used in the *T'ai Chi Teacher*. 13

3-1 Real-time estimation of the position of moving hands and head 24

3-2 Hypothetical PDF's for an "Opening Form" gesture in (x, y, z) 25

3-3 Hypothetical PDF's for "Opening Form" in (dx, dy, dz) 26

3-4 An example HMM topology illustrating skip states 27

3-5 The topology and transition probabilities for the "Single Whip" model. . . 28

3-6 The learned PDF's for "Single Whip." 28

4-1 Time windows fed to Viterbi algorithm 31

5-1 Probability along the observation sequence 37

5-2 Calculation of a sample error percentage 37

5-3 Probability tracks and thresholds for the gestures 39

B-1 An example of the raw and resampled/filtered data. 46

List of Tables

5.1 Recognition accuracy in offline tests 34

5.2 Confusion matrix for real-time recognition of multiple users 35

5.3 Results of the experiments testing the identification of salient moments. . . 38

Chapter 1

Introduction

1.1 Statement of Thesis

The real-time, interactive feedback system we developed, *Sensei: The T'ai Chi Teacher*, is a novel use of virtual reality: a teaching tool for *T'ai Chi* gestures. Furthermore, it provides a platform on which to build a more sophisticated teaching, training, and feedback tool for gestures or action.

In this document, we hope to substantiate this thesis by showing that *Sensei* does have the components of a good foundation. First of all, it is capable of performing real-time, multiple user, gesture recognition. Users of the system are free to practice *T'ai Chi* gestures, starting anywhere in the sequence of moves, and the system recognizes their actions. In addition, a complete teaching system must be able to give both positive and critical feedback to the user. This ability implies a knowledge of the instants in the user's gesture where the user was both least and most accurate in her movements[1].

1.2 Motivation

The system described in this document, *Sensei: The T'ai Chi Teacher* is one part of a larger application called *Staying Alive.* A complete discussion of that applications falls outside the scope of this thesis; however, as a general context and as motivation, it is necessary, and so will be discussed here and in a later chapter directly about related work.

[1] For lack of a better term, we subsequently refer to these moments as the salient moments in the gesture.

1.2.1 Staying Alive

Growing evidence supports the idea that mental imagery can have a profound healing effect on the human body. For example, numerous studies have shown that from imagery springs forth directly measurable effects such as the cure of warts or the reduction of scarring from burns [17, 37]. Furthermore, several studies have shown the efficacy of programs involving imagery and relaxation in increasing the expected lifespans of cancer patients [24, 34]. While it is difficult to prove which particular aspects of such programs (imagery, relaxation, expectation, or group meetings) account for the changes in prognoses, it is reasonable to assume that imagery has some effect, given the types of studies described above.

While the use of imagery has been shown to directly correlate with healing, the "relaxation response" associated with relaxation methods has been shown to boost the immune system's activity in general [8]. Stress seemingly shuts down the immune system, perhaps to prepare for a "fight or flight" response to a perceived threat; relaxation allows the immune system to flourish. For this reason, relaxation methods are commonly taught to cancer patients. In addition to helping the besieged immune system, relaxation methods can help cancer patients feel a sense of control at a time in their lives when cancer treatments cause inordinate stress and anxiety. This sense of control itself, even if not directly correlated to prognosis, certainly improves the quality of life.

Because imagery and relaxation promise so much benefit to both those with and without cancer, Becker and Pentland are developing a virtual reality imagery and relaxation tool called *Staying Alive* [7]. The long term goal of that project is to determine whether using virtual reality as an imagery device is more beneficial than using the imagination alone. In order to create the imagery, Becker and Pentland have developed a virtual environment of a bloodstream. In *Staying Alive*, the user controls a white blood cell in the environment and navigates through the blood stream, removing malignant cells. This immune system-centered paradigm is a common imagery theme for cancer patients.

The other aspect of *Staying Alive*, relaxation, is currently in development. The goal is to allow a user to control the virtual environment through the use of *T'ai Chi* gestures. *T'ai Chi* is among the more popular forms of martial arts in China and the world. Though like all martial arts, this exercise includes movements for attack and defense, there are several forms of *T'ai Chi* which are slow and gentle, such as the Yang, Wu or Sun styles [41]. The gestures for *Staying Alive* were chosen from the Yang style, which is meant for relaxation and wholeness, not for fighting. *T'ai Chi*, like imagery techniques, is commonly taught at cancer centers such as the Wellness Community both as a gentle way to exercise and as a relaxation method. Its use in *Staying Alive*, then, is two-fold: both as a relaxation tool and as the means by which to control the environment.

1.2.2 Sensei: the T'ai Chi Teacher

An application with *T'ai Chi* gestures as part of the user interface is only usable by those who are familiar with *T'ai Chi.* Therefore, as part of the development of the *Staying Alive* application, we were motivated to produce an application that could teach users the gestures necessary to control the environment. The goal of this work centers directly on that teaching application. Specifically, the long-term goal is to develop a system capable of providing users with feedback and training as they practice *T'ai Chi.*

1.3 Approach

Embedded in the task of developing a system capable of giving useful training are two other issues: the system must be capable of recognizing *T'ai Chi* gestures when performed and then of comparing the user's movements to an ideal, or perfectly acted, gesture to provide feedback.

The *Sensei* system tackles the first problem, that of gesture recognition, through a Hidden Markov Model approach. The user's head and hands are tracked in 3-D via *STIVE* (Stereo Interactive Video Environment) [4]. Cartesian (dx, dy, dz) velocities are used as features, and and left-right, Bakis HMM's [26] for each gesture are trained with the Baum-Welch algorithm. Recognition is achieved in real-time with the use of the Viterbi algorithm. The five gestures used in the system can be seen in Figure 1-1. These five gestures were chosen because they are the first five gestures in the short, Yang *T'ai Chi* sequence [41].

The probabilistic framework of Hidden Markov Models generalizes well for computing the moment in the user's expression of a gesture that differs most from the ideal form. In the course of using the Viterbi algorithm to compute the probability of each model, the probability per frame during the course of the gesture is computed. This can be used not only to do recognition, but also to identify the moment in the gesture sequence where the user most differed from (or most approached) an expert.

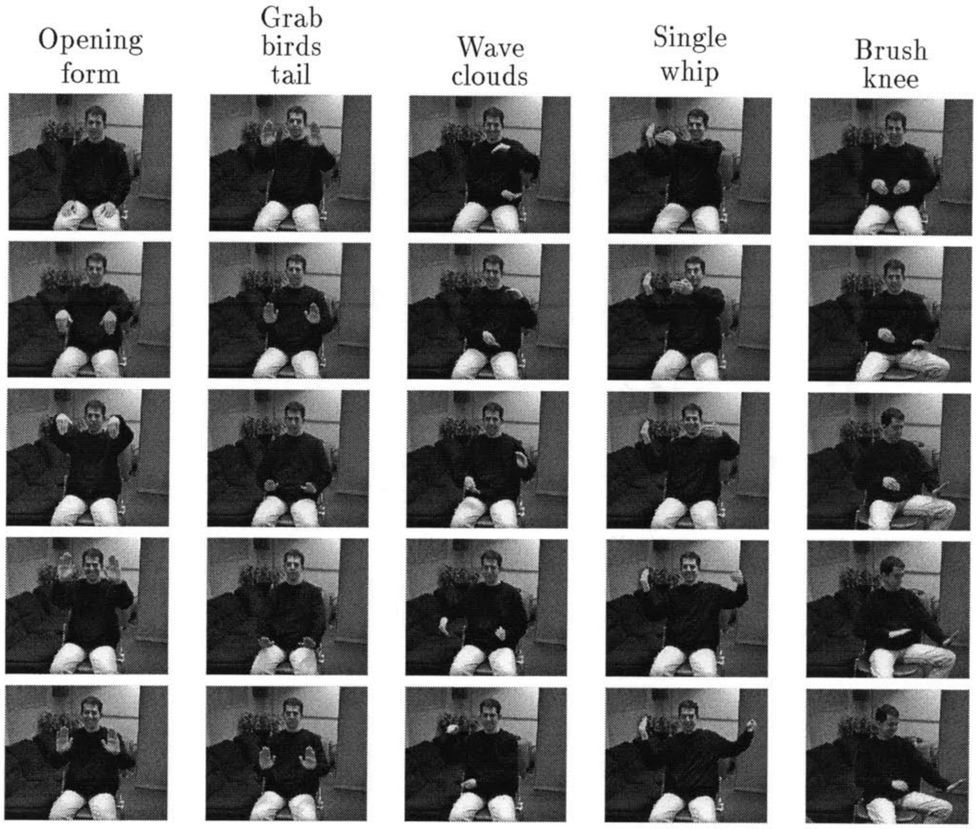


Figure 1-1: The five *T'ai Chi* gestures used in the *T'ai Chi Teacher*.

1.4 Other Applications

Imagery techniques are currently used in many applications. Imagery and hypnosis (which is a focused way of doing imagery) are commonly used in such widely varying arenas as pain control and athletics. An example imagery for someone coping with pain is to picture a sailboat floating away on the ocean and to imagine the sailboat carrying away the pain as it disappears on the horizon. An athlete might try to improve her free throw shooting ability by picturing herself successfully making her shots. Little has helped both cancer patients and several professional athletes with programs of imagery and relaxation [20].

Not all imagery systems would benefit from a gesture recognition interface. However, as an application in itself, a system that watches a person perform some physical activity and acts as a teacher could be applied in many areas. For example, such a system could act as a tennis instructor, a golf pro, an aerobics instructor, or a dance teacher. In general, the system presented in this document could be expanded into a general framework for teaching physical movements of any kind. An expert would teach the system the correct way to perform the movements. Users would have their movements recognized by the system and then receive feedback as to where they can improve.

1.5 Outline of Thesis

The remainder of this document explores the details of our approach to building *Sensei: The T'ai Chi Teacher.* Chapter 2 provides the context of the work in relationship to previous work. Chapter 3 discusses the implementation and training of the gesture recognition system. Chapter 4 contains a description of the recognition algorithm and the feedback system. Experiments assessing the effectiveness of the system and the results of those experiments are presented in Chapter 5. The final chapter is a summary, with a brief look at future work.

Chapter 2

Related Work

The ideas motivating the *Sensei* system were inspired by work in several fields. Research about the use of imagery and relaxation to bolster the immune system has occupied a role in psychosomatic medicine for years. This work directly inspired the *Staying Alive* application, of which the *T'ai Chi Teacher* is a part. The first section of this chapter will discuss some of that work. The field of gesture recognition is a rapidly growing subset of computer vision; some of the works inspiring the Hidden Markov Model approach used in *Sensei* will be discussed in the second section.

2.1 Imagery and Relaxation

2.1.1 Early Work in Imagery, Visualization, and Hypnosis

For centuries, Tibetan monks have been using elaborate visualizations and meditations to control parts of their physiology which are normally under autonomic control. For example, Dr. Herbert Benson, a Harvard cardiologist, recorded monks raising their skin temperature by as much as seventeen degrees above normal while sitting in near-freezing temperatures, wrapped in wet sheets. "If an ordinary person were to try this, they would shiver uncontrollably and perhaps even die. But here, within three to five minutes, the sheets started to steam and within forty-five minutes were completely dry," says Dr. Benson [8].

Benson termed this physiological response to this meditative state the Relaxation Response. Essentially, when a person perceives stress, the brain enters into a "flight or fight" response and releases hormones into the bloodstream to prepare the body. These hormones are known to have a suppressing effect on the immune system. The Relaxation Response, on the other hand, triggers the exact opposite reaction. The release of those hormones is quelled and the immune system functions at its full capacity.

Can people use a similar technique to encourage the body to heal itself, rather than raise the skin temperature? The most medically unequivocal cases of the mind healing the body come from studies of warts. Warts, which are caused by a virus, appear to be unusually susceptible to hypnotic suggestion. The imagery and meditations used by the monks are similar to the hypnotic inductions used to treat warts. Both typically involve a relaxation phase and a guided imagery. As early as 1927, the first scientific studies on this issue were carried out by Block [9]. He used suggestion without hypnosis in his work, blindfolding his patients and painting their warts with an inert dye. More than half his patients lost their warts within three months (untreated, warts resolve spontaneously in 2.28 years on average [37]). Memmesheimer and Eisenlohr followed this study with a more controlled experiment showing a similar connection between the cure of warts and suggestion [22]. In the 1950's, another series of studies probed the issue of hypnotizability and warts [21, 12, 2, 3, 38]. Though these studies focused on the relationship between hypnotizability and efficacy of hypnotherapy, the idea that warts were susceptible to hypnotic suggestion (which usually entails extensive use of imagery) was reinforced. In 1972, Surman, *et. al.*, carried out a rigorously controlled study testing specifically whether warts were susceptible to hypnotherapy. While the warts of 53% of their experimental group improved, not a single member of the control group experienced any improvement. Surman concludes that while hypnosis is clearly an effective treatment, one can only conjecture as to *how* hypnosis encourages the mind to cure the body of warts.

2.1.2 Imagery as Applied to Cancer

Work such as that described above clearly established the mind's capability of influencing the health of the body. Carl Simonton, *et. al.*, was the originator of the idea of using these techniques with cancer patients [30]. If imagery and relaxation had been effective in giving control over parts of the body previously considered outside conscious control, they might give the body a way to focus the attention of the immune system on the cancer.

Several researchers have investigated whether Simonton's hypothesis is correct. Simonton reported a life span extension of over 60% for patients enrolled at his center during its first four years [30]. Achterburg and Lawlis also studied a group of patients enrolled in the Simonton's program and found consistently better results for patients that practiced imagery daily [1]. Achterburg argues, in fact, that the imagery itself is the important therapeutic component in the healing program.

In a broader study, Spiegel, *et. al.*, demonstrated that psychosocial interventions (which included group therapy and self-hypnosis) almost doubled survival time in the experimental group compared to the control group [34]. Several other researchers have shown relaxation and imagery to be an effective means of reducing stress and anxiety in cancer patients, without studying the specific effect on long-term survival [28, 15, 14, 18, 10, 31, 6]. Postwhite has not only shown the positive effects of imagery on emotion, immune function, and cancer outcome, but has shown that it causes an increase in lymphokine activated killer cells [24].

While the community studying the effect of imagery and relaxation on cancer patients has demonstrated a correlation between the practice of such measures and outcome, the hypnosis community has continued to refine its work on the treatment of warts. Spanos, *et. al.*, have shown that in their studies, the subjects who had the most success in curing warts were those who reported having the *most vivid imagery* as opposed to those who were most hypnotizable [32]. In fact, they found that hypnotizability did not correlate with results at all. In a later study, Spanos, *et. al.*, confirms again that vividness of imagery facilities wart loss. They further hypothesize that it is the enhanced sense of cognitive involvement in those patients that produce the most vivid imagery that is the true underlying cause of the increased wart loss [33].

Finally, Baer and Surman demonstrated as early as 1985 that computers were effective tools in inducing relaxation and focused attention. In their study, twenty adults used an APPLE IIc program and rated their anxiety via the Spielberger State Anxiety Scale. They found that the computer was an adequate tool for assisting adults in relaxation and stress-reduction [5].

In summary, there is a wide range of work showing that imagery and relaxation are beneficial to the function of the immune system. Furthermore, the vividness of the imagery experienced by the practitioner correlates most strongly with how effective hypnosis and imagery are in curing warts. Lastly, Baer and Surman demonstrated the feasibility of using a computer to enhance relaxation. These issues all suggest that a virtual reality application, in which the vividness of imagery is directly controlled, might be an effective tool against cancer. That is the primary experimental question which *Staying Alive* attempts to answer.

2.2 Using HMM's for Gesture Recognition

The continuous speech recognition community has embraced the use of Hidden Markov Models (HMM's) for years [27, 26, 19]. The ability of HMM's to use dynamic time warping to provide time scale invariance while maintaining a probabilistic framework has more recently made them attractive to the computer vision community. In addition, their ability to automate segmentation and classification makes them well suited for gesture recognition.

Early work with HMM's in vision was done by He and Kundu [16], who used them to classify planar shapes. Their work derives more closely from work in the handwriting recognition community.

Yamato, *et. al.*, used HMM's to recognize three different subjects performing six different tennis swings. As input features to the HMM's, they used a 25 × 25 quantized, subsampled image. With thirty instances for each stroke used to train the models, successful classification between them was achieved [40].

Schlenzig, *et. al.* [29], demonstrated the ability of HMM's to recognize continuous gestures from image sequences, rather than from still frames. As an input feature vector, Schlenzig uses a rotation invariant, binarized frame around the hand, processed by a neural net. Their system is capable of distinguishing between "hello", "good-bye", and "rotate".

The recognition system of *Sensei: The T'ai Chi Teacher* is most directly influenced by work done by Starner and Pentland [35]. Using hand velocities and orientations in two dimensions, Starner was able to build an HMM system capable of recognizing forty American Sign Language gestures in a real-time system. The features are computed using a system that tracks hands wearing colored gloves. The system is capable of recognizing gestures with an accuracy of 97%.

Wilson and Bobick [39] develop a state-based method of learning visual behavior of gestures in an image sequence. Multiple representations are fed into an HMM and the input's overall membership in a given state is determined by which representation best describes the input.

Darrell and Pentland have explored a real-time wireless hand gesture recognition system that does not use HMM's directly. Instead, their work is a view-based approach in which matched filters are acquired for examples of different gestures and where a new filter is learned whenever an example is displayed for which no previous filter is well suited [13]. The relationship to the *T'ai Chi Teacher* is in their use of dynamic time warping to assist in recognition. The use of the Viterbi algorithm to achieve recognition in the *T'ai Chi Teacher* uses a similar dynamic programming technique.

Finally, the gesture recognition system of the *Sensei* system was derived directly from work done by Campbell, Becker, Azarbayejani, Bobick and Pentland [11]. In this work, we test a set of seven different feature vectors, all of were which were functions of coordinates of the hands and the head of a user, as input to an HMM system. The positions of the hands and head are tracked in real-time in three dimensions with the use of a wide-baseline stereo camera system. Our results showed that with invariances to user translation and rotation as a goal, polar, body-centered velocity coordinates can achieve 93% recognition accuracy on a vocabulary of 18 different *T'ai Chi* gestures.

The main difference between this previous work and the work done for the *T'ai Chi Teacher* is in the recognition, rather than training, phase. In the earlier work, we hand-segment gesture sequences into groups of six, and entire sequences are parsed at once. The recognition routine utilizes the grammar of there being exactly six gestures in the sequence when doing recognition. The desire to allow a user to practice a sequence of gestures, starting at any point, and to give feedback in real-time necessitated a system that does no segmentation at all. Instead, the *Sensei* system observes the user in real-time and performs both gesture recognition and gesture spotting.

Chapter 3

Implementation of the Recognition System

The previous chapter discussed several systems which utilized Hidden Markov Models to perform gesture recognition. The work done by Starner in particular, in which significant recognition success was achieved with a vocabulary of forty American Sign Language gestures, suggests that this framework is appropriate for the recognition task of the *Sensei* system [35]. In addition, and perhaps more importantly, Hidden Markov Models have the attractive feature of placing the entire recognition task in a probabilistic framework. This framework allows the *T'ai Chi Teacher* to easily pick out segments of the user's gesture which most closely and least closely fit to the ideal version of the gesture, an aspect that will be discussed in more detail in the next chapter.

3.1 Description of Hidden Markov Models

Hidden Markov Models are based on the assumption that the process being modeled can be described as a first-order Markov process. Such a process is one in which the system can be expected to jump from state to state over time. The system's parameters at any given time are described by the state in which the system currently resides. As the system changes in time, its parameters might change and be better described by a different state. The change from one state to another is a stochastic process. The Markovian property states that for a first-order Markov process, the probabilities of transitions between states depend only on the current state (a second-order Markov process would be one in which the transitions between states depend only on the last *two* states, *etc.*).

The difference between a first-order Markov process and a Hidden Markov Model is that in the HMM framework, the current state of the system is not observable. Instead, the system outputs a symbol at each time step, where the symbol is generated stochastically by whichever (unobservable) state in which the system currently resides. That is, there is a stochastic process governing which state the system is in and another stochastic process, characteristic of the current state, which determines which observable symbol the system outputs. The system, then, is doubly stochastic.

As such, a Hidden Markov Model with n states can be described completely by the following quantities:

- initial probabilities, π: an n-vector where π_i is the probability of starting in state i
- transition probabilities, A: an $n \times n$ matrix where a_{ij} is the probability of jumping from state i to state j
- output probabilities, B: HMM's can be either discrete or continuous. Discrete HMM's have output that is one of m possible symbols. In this case, B is an $n \times m$ matrix, and b_{jk} describes the probability of state j outputting symbol k. In the continuous case, the observable symbol output by the system is a continuous random vector. B now describes parameters for a set of probability density functions (typically a mixture of Gaussians) which give probabilities for different observable vectors.

Thus, an HMM can be fully described by $\lambda = (\pi, A, B)$.

3.2 Implementation of HMM's

3.2.1 Topology

To use HMM's as a recognition tool, then, several steps must be taken. First of all, each HMM must be given a topology. In general, knowledge of the physical properties of the system to be modeled can be used to create an appropriate topology. For example, consider a traffic light that during the day outputs either green, yellow, or red. At night, the light enters a different mode in which it outputs a blinking yellow. An HMM that models this traffic light might, then, have two states, each with a quite different output probability function.

In practice, of course, real systems are rarely as simple. In general when designing a system, one starts with a topology which is suspected to be more complicated (*i.e.* has more states and more paths between the states) than the system being modeled. While training the model, it is then possible to prune the topology by removing states and links which are seldom used. Stolcke and Omohundro have attempted to automate this process. See [36] for details.

3.2.2 Choosing the observation vectors

Another question to be answered when developing an HMM system is the choice of what to use as the observation vector. Consider the traffic light system again. The observation is of the color lit by the traffic light. In this case, the HMM would be a discrete system; the observations are simply one of three different colors. The state remains hidden in that there is no directly observable quantity which immediately specifies the state. Only by analyzing the pattern of an observation sequence can the state be deduced. In practice, of course, it would be simple to discern which state the light was in by observing the pattern of lights and noting the presence or absence of the repeating yellow. This step is exactly what the HMM recognition system will eventually do – determine the most likely state from the observation sequence.

The choice of observation vectors when implementing real systems is affected by issues such as available sensors, desired invariances (*e.g.* speaker independence, view independence, *etc.*), and amount of available training.

3.2.3 Training the model

The next task to accomplish is training the system, using example data to learn appropriate transition and output probabilities. The goal is to take an observation sequence known to have come from a certain model and change λ such that the probability that the given model produced the observation sequence is maximized. In general, this is accomplished through the use of the Baum-Welch algorithm. This algorithm is an iterative re-estimation routine, guaranteed to find a local maximum of the probability. While the probability surface is likely to be quite complex, in practice, the Baum-Welch algorithm is effective at quickly arriving at adequate models.

3.2.4 Performing the recognition

Finally, once models have been trained for all of the atoms to be identified (*i.e.* gestures in a gesture recognition system, words in a speech recognition system), the Viterbi algorithm can be used to perform recognition. The Viterbi algorithm is based on dynamic programming techniques and bears close resemblance to dynamic time warping. The task of recognition in the HMM framework is to take a given observation sequence and determine which of the HMM's was most likely to have emitted it.

The procedure works by maintaining a lattice structure of probabilities. Each column in the lattice, δ_i, has n nodes, each of which represents the probability of being in a given state. There are as many columns in the lattice as there are observations in the sequence. The lattice gets filled in recursively, starting with the first observation. The initial nodes are given a probability of:

$$\delta_i = \pi_i b_i(O_1)$$

Then, nodes at time t are filled in with:

$$\delta_t(j) = Max_i[\delta_{t-1}(i)a_{ij}]b_j(O_t)$$

The lattice gets filled in until the last observation, at which point the node with the maximum final probability is chosen, and the sequence can be recovered by back-tracing through the lattice[1].

3.3 Implementing the T'ai Chi Teacher's HMM's

3.3.1 Observation vectors

We now turn to a description of the particular implementation used in developing *Sensei*. The first implementation question is the choice of what to use as the observation vector. Starner's American Sign Language recognition system was based on the 2-D mean coordi-

[1] This method chooses the most likely state sequence for the observation data. The procedure can be modified to maximize the most likely state at each time step, as well. Typically the choice of the appropriate version is predicated by the problem. Ergodic models, in which every state is reachable from every other state (*i.e.* $a_{ij} > 0$ for all i,j) use the latter version. Bakis, left-right models (*i.e.* $a_{ij} = 0$ for $j < i$) use the former to enforce that the final sequence chosen is valid.

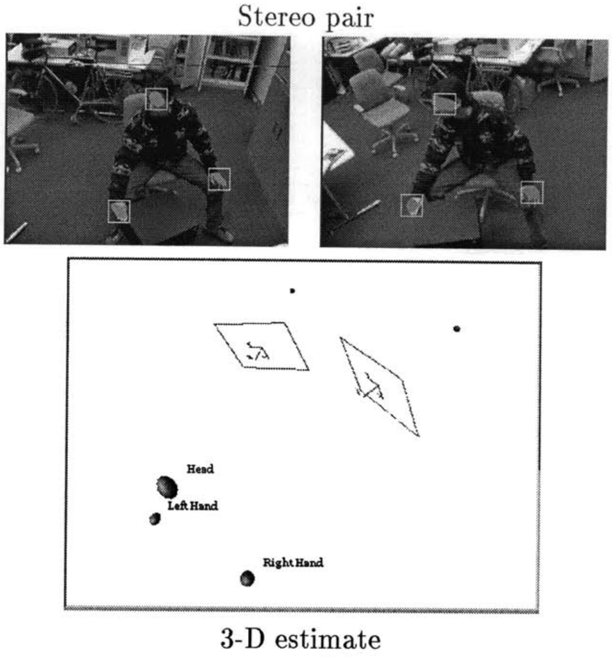


Figure 3-1: Real-time estimation of the position of moving human head and hands. The top images are the video and tracking from the two cameras and the bottom image is the result of triangulating.

nates of both hands, as well as a description of the hands' orientations [35]. For the *T'ai Chi Teacher*, we felt that the use of functions based on the 3-D coordinates of the hands and head would be more appropriate. First of all, in previous work, we have demonstrated the effectiveness of several different such functions, using *T'ai Chi* gestures [11]. Perhaps more importantly, to build a platform for a more general teaching tool, one which is capable of distinguishing between gestures from an arbitrarily complex vocabulary, it is reasonable to assume that 3-D data might be necessary. The *T'ai Chi Teacher*, as a testbed for future, more sophisticated teachers, must demonstrate the ability to recognize complicated gestures using the 3-D data.

To gather this 3-D tracking data for the hands and head, we utilize the *STIVE* (Stereo Interactive Video Environment) system developed by Azarbayejani and Pentland. In the smart desk environment of *STIVE*, two wide-baseline cameras are positioned at the top of a large display screen in front of which sits the user. The video from each camera is separately analyzed to find two-dimensional blob features [4]. Essentially, this task is accomplished by searching the video input stream for flesh chrominance and then using a geometry model to decide which blob corresponds to the left hand, the right hand, and the head. Once these labeled, two-dimensional blobs are known (along with the calibration of the cameras), a

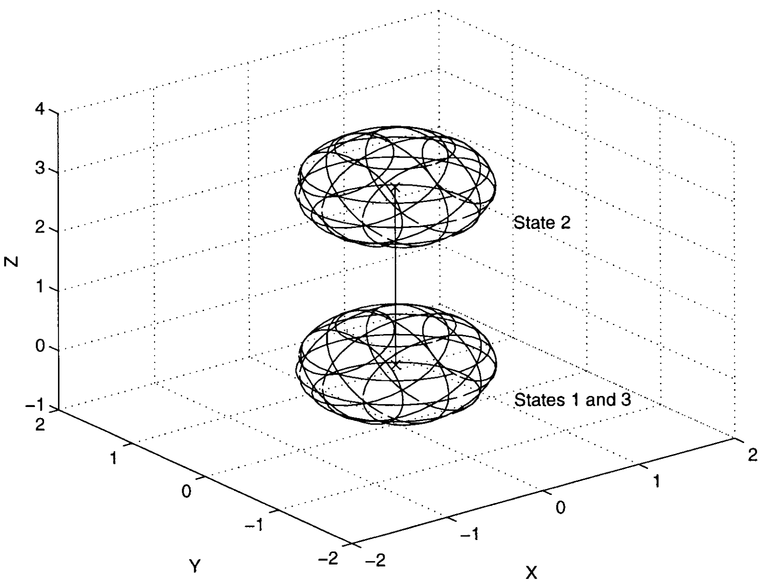


Figure 3-2: Hypothetical PDF's for an "Opening Form" gesture in (x, y, z). Notice that states 1 and 3 overlap.

recursive, nonlinear estimation is used to extract an estimate of the three-dimensional blob features of the hands and head as in Figure 3-1. *STIVE* is capable of tracking head and hand positions with an accuracy of approximately 2cm at almost 30 frames/sec. Although the system estimates the full blob features (which include orientation as well as location), for the purposes of the *Sensei* system, we use the mean location of the blobs as the locations of the hands and heads.

One possibility for the choice of observation vectors would be to simply use the position data as provided by *STIVE*. The models learned in that case, would have states corresponding to the presence of the hands and head in a specific area of space. For example, consider the *T'ai Chi* gesture, "Opening Form". This gesture (see Figure 1-1) consists of raising both hands up, holding briefly, and then bringing both hands back down. If positions are used as feature vectors, the states of the model would correspond to probability density functions in (x, y, z) along the trajectory of the hands, as in Figure 3-2.

This choice of features, however, has limitations. If a new user performs a gesture without being in the exact same location as the training data, it will prove to be a poor match because there is no invariance to user translation. To avoid this problem, we use velocity features as our observation vectors. In this case, the probability density functions of the HMM states are not tied into specific locations in space, but instead to magnitudes of velocity. For "Opening Form", this means that the model would have state(s) representing

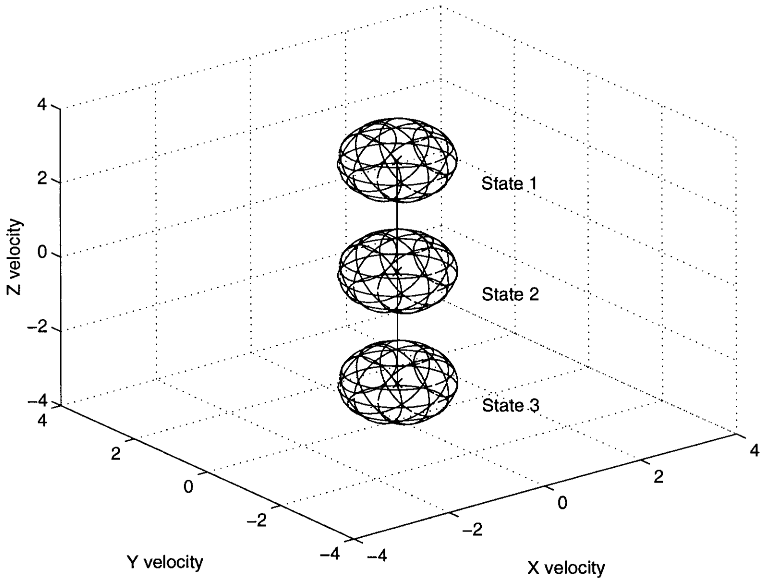


Figure 3-3: Hypothetical PDF's for an "Opening Form" gesture in velocity space.

an initial positive vertical velocity, then a state(s) representing a rest at the zenith, and then state(s) representing a negative vertical velocity as in Figure 3-3. This set of features is shift invariant; users in different locations performing the same gesture would have the same data. It is not rotation invariant, however; users which are rotated relative to one another will have different non-vertical velocities. In the case of the *T'ai Chi Teacher*, we can assume that the user is facing the view screen, so rotation invariance is not a goal. For a more complete discussion of the issues surrounding feature vector selection, see [11].

3.3.2 Choosing a Topology for the T'ai Chi Teacher

The motion of *T'ai Chi* gestures proceeds in a time-ordered manner. That is, gestures proceed from a beginning to an end. To model signals with that property, a Bakis model in which

$$\pi_i = \begin{cases} 1 & i = 1 \\ 0 & i \neq 1 \end{cases}$$

and $a_{ij} = 0$ for $j < i$ (so transitions may only proceed in a forward manner through the states) is most appropriate. The next issue is to define the number of states for each model. What if all the gestures are not equally complicated, however? Perhaps some would be suitably described with a fewer number of states than others. In order to avoid developing

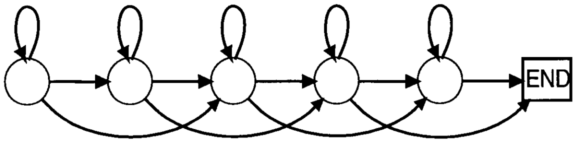


Figure 3-4: An example HMM topology illustrating skip states

a different model for each gesture, skip transitions may be utilized, as in Figure 3-4. These transitions allow the same topology to accommodate a different effective number of states.

To produce the actual topologies used in *Sensei*, we considered the physical properties of the five gestures. Three of the gestures, "Opening Form", "Single Whip," and "Brush Knee" all consist grossly of three different movement segments. For example, consider the gesture "Opening Form." This gesture begins with an upward motion of the hands, then a pause while the palms are brought from a horizontal to a vertical position, and then finally a drop of the hands. "Grab the Bird's Tail" and "Wave Hands Like Clouds", on the other hand, have more motion segments.

Given this physical scenario, we began by using a five state model with three skip states, as in Figure 3-4, effectively allowing the training to generate models with three, four, or five states. After completing the training, we found that the three simpler gestures all had high skip probabilities. Therefore, we pruned their models down to three states with no skip states, leaving the other models as is, and retrained.

3.3.3 Training the Models

In order for the system to be capable of recognizing multiple users, it is necessary to train the system on a variety of experts. This way, when learning the models, the training data will contain the types of variations that occur across different performers. To accomplish this goal, 15 examples of each gesture were collected from four different performers in a random sequence of five moves at a time, for a total of 60 examples of each gesture. A commercial program, Hidden Markov Model Toolkit, by Entropics, was used to run the Baum-Welch training algorithm.

As a result, five different HMM's were developed to represent the five different *T'ai Chi* gestures to be recognized. To understand the physical meaning of these models, consider the model for the gesture "Single Whip". In order to visualize this model, we will consider

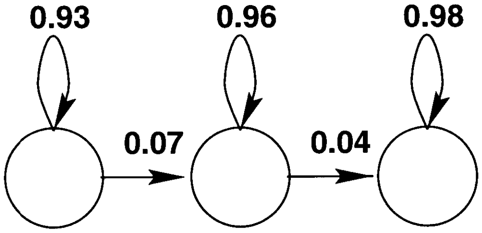


Figure 3-5: The topology and transition probabilities for the "Single Whip" model.

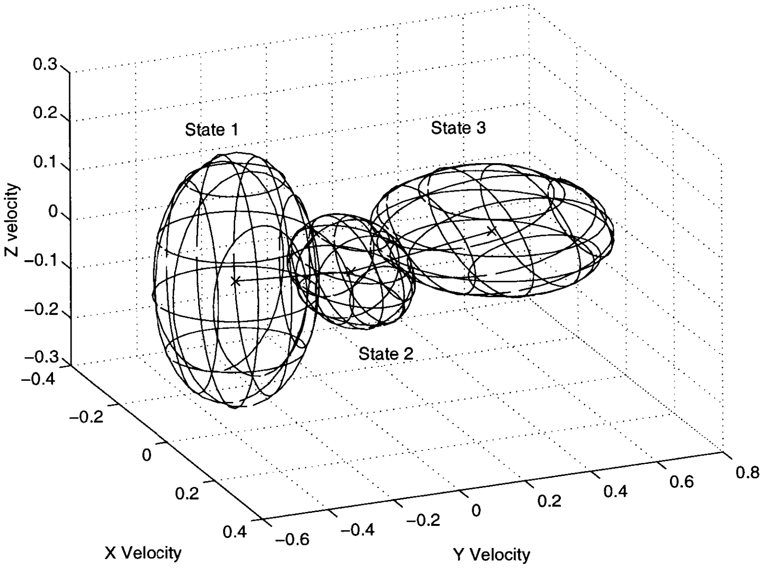


Figure 3-6: The PDF's for "Single Whip". The sizes of the axes of the blobs correspond to the variances in those directions.

only the motion of the right hand (the left hand is motionless in this gesture, anyway)[2]. The right hand starts to the left of the head, held slightly out from the chest. Next, the hand moves across the chest to the right, pauses, and returns. In velocity space, this is a negative horizontal velocity in y, followed by a pause with no velocity, and completed with a positive horizontal velocity in y. Figure 3-5 shows the topology for the model with transition probabilities, while Figure 3-6 shows the probability density functions for each of the states. In general, then, the five models learned during training represent the gestures by these blobs in velocity space and the transition probabilities.

[2] We do not use full covariance matrices for the output probabilities, so it is valid to separate the probability density functions into separate 3-D blobs for left and right hands.

Chapter 4

Recognition of Gestures and Feedback

4.1 Gesture Recognition

In the previous chapter, we briefly discussed how the Viterbi algorithm takes a Hidden Markov Model and an observation sequence and determines the probability that the model produced the sequence, $Pr(\mathbf{O}|\lambda)$. In order to use the Viterbi algorithm to perform recognition in real-time, however, two obstacles must be overcome.

- **Segmentation:** The previous discussion of the Viterbi algorithm assumed that the data was previously segmented. That is, the algorithm returns a probability for the most likely state sequence traversing the entire observation sequence. In the real-time environment of the *T'ai Chi Teacher*, the goal is to track a user doing a series of *T'ai Chi* gestures and recognize whenever they have completed any of the gestures.

  To accomplish this goal, we check for the presence of the gesture in the observation sequence ending with the current frame, but beginning anywhere over a range of times in the past, as in Figure 4-1. In this way, we hope to be able to find gestures which can be of varying lengths, either due to inherent differences in their structures or due to user variability. We define two parameters: a minimum time allowed for a gesture, T_{min}, and a maximum time allowed for a gesture, T_{max}. Frames of tracking data are stored until they are older than T_{max}, at which point they are discarded. With each new frame, the Viterbi algorithm is run for each gesture on all the observation

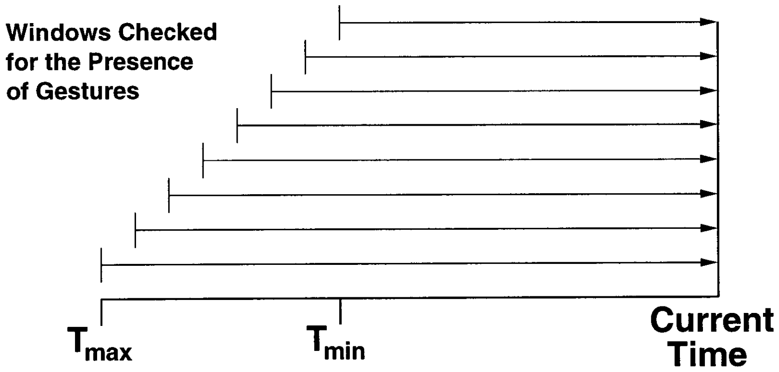


Figure 4-1: With time proceeding to the right, this plot illustrates the numerous windows fed to the Viterbi algorithm. The lengths of the windows varies from T_{max} to T_{min}.

sequences starting from T_{max} through T_{min} and ending with the current frame[1]. The probability for each gesture is then

$$Pr(\lambda) = Max_{T_{min}<t<T_{max}}[Pr(\mathbf{O_t}|\lambda)]$$

In order to compare probabilities between sequences of differing lengths, the average probability per frame is used.

- **Thresholding:** The other issue which the real-time constraint imposes is the issue of thresholding the probabilities. When data is presegmented into gesture sequences, recognition is simply a matter of determining the maximum probability of the observation sequence over the models. When tracking a user in real-time, who may or may not be doing a gesture at any moment, it is necessary to redefine the "recognition" of a gesture to be only a moment in time when the probability for a certain gesture exceeds some threshold value. In this way, the gesture recognition acts as a "word spotting" tool as well. The value of the threshold for the *T'ai Chi Teacher* was derived empirically for each gesture by experimentation.

[1] Actually, it is computationally expensive to change the beginning frame in the Viterbi probability lattice. To avoid this problem, we time reverse the data and effectively vary the ending frame of the observation. See Appendix A for details.

4.2 Localizing Salient Moments in the Gesture

In order for a more sophisticated teacher to give appropriate feedback to a user, it must be able to locate points in the user's actions where the gesture was particularly "good" or "bad." The final task of the *Sensei* system, then, is to analyze an observation sequence which has been recognized as a particular gesture and identify these points. Later, fully developed teachers can then use this knowledge to provide feedback.

Before building a tool to locate these salient moments, we must first define what makes such a moment. To ground the definition, we consider the user's motion in relationship to an "ideal" version of the gesture. To develop this ideal version of each gesture, we use the training data used to build the HMM's, which was gathered by tracking experts performing the gestures. To develop the ideal gestures, these expert examples are linearly resampled in time and averaged.

Given the ideal versions of the gestures, the simplest definition of the salient moments in the user's action is to consider the points in the user's motion where the hands were closest or furthest (in a Mahalanobis distance sense) from where the hands were in the corresponding time in the ideal gesture. Among the drawbacks to this definition are that this leaves no shift invariance – a user must perform the gesture in the same position relative to the tracking system as the experts did – and it has no size invariance – people of different height and arm length must move their hands through the exact points in space that the experts did.

A more suitable definition that does provide shift and position invariance is to consider the same distance metric, difference from the ideal version of the gesture, but in velocity (dx, dy, dz) space. This means that gestures are compared on the basis of whether they involved movements in the correct directions and magnitudes as compared to the ideal version.

Furthermore, there should be some flexibility in the rate at which users proceed through the gesture. That is, it might take some people less time to complete a motion than others because of different heights or arm lengths. In general, in *T'ai Chi*, the speed of the motion should be consistent across practitioners – different users should not move their hands at different speeds through space. However, it might take more or less time for different people to move their hands through the proper trajectories *while traveling at the same speed.* For larger people, the trajectories are simply longer in space. In summary, then, according to this definition, salient moments in the gesture occur when a user was most and least aligned to the ideal version of the gesture in velocity space, with some flexibility for how long each motion segment is.

The Viterbi algorithm, already being employed in the *Sensei* system to perform recognition, can also be used to locate these salient moments in the user's gesture. Once recognition has been achieved, the observation sequence has been successfully segmented. In computing the probability lattice described in the previous chapter, the Viterbi algorithm computes $Pr(\mathbf{O}|\lambda)$ at each point in the observation sequence. In effect, the observation sequence is mapped in velocity (dx, dy, dz) space to the probability density functions of the states of the HMM. If the hands move at a higher or lower speed than the experts, the observations will fall to the edges of the PDF's. However, the dynamic time warping capability of the Viterbi algorithm allows the hands to move at the proper speed for a variable period, without penalty. In other words, the Viterbi algorithm computes the probability that the user's gesture matches the ideal gesture at each point in the gesture sequence, while allowing for flexibility in the correspondence. To pick the salient moments in the gesture, all that is required is to pick the minimum and maximum probability along the sequence.

Chapter 5

Experiments

5.1 Gesture Recognition Experiment

Our previous work demonstrated that velocity features of 3-D tracking data of the hands can be used to achieve a highly accurate recognition rate [11]. In that system, 18 examples of 18 different *T'ai Chi* gestures were used as training, and testing was performed on 6 examples of each of the 18 gestures. No portion of the testing data was used during training. The testing sequences were hand segmented into groups of six, and this knowledge was utilized as a grammar by the recognizer. Therefore there was no chance of insertion or deletion errors, only substitution errors. A recognition rate of 98% was achieved on testing examples recorded with the user in the same position and orientation as in the training examples.

Given this work, it is clear that Hidden Markov Models are an effective tool for recognizing pre-segmented gestures. However, the recognition task of the *T'ai Chi Teacher* is slightly different. This application must be able to recognize gestures from an unsegmented stream. In addition, the previous work used the same person for testing and training data. As a general teaching tool, *Sensei* must be able to recognize different people performing *T'ai Chi* gestures.

To first test the plausibility of the models, recognition was performed offline on recorded

training set	*independent test set*
97.8%	95.1%

Table 5.1: Recognition accuracy in offline tests

	OF	GBT	WC	SW	BK	Accuracy
OF	55	5	0	0	0	91.6%
GBT	2	57	1	0	0	95.0%
WC	1	2	56	0	1	95.0%
SW	0	0	0	59	1	98.3%
BK	0	0	0	1	59	98.3%
Total						95.3%

Table 5.2: Confusion matrix for real-time recognition of multiple users. None of the users were in the training data. OF:Opening Form, GBT:Grab Birds Tail, WC:Wave Clouds, SW:Single Whip, BK:Brush Knee. Row labels are true nature. Column labels are system classification.

data. Four users were recorded performing 15 examples of each gesture, in sentences containing all 5 gestures in random order. In the first experiment, all 60 examples of each gesture were used to train the models and then all 60 examples were used in testing. In the second experiment, training was done on 45 randomly selected examples of each gesture and testing was done on the other 15 examples. Table 5.1 shows the recognition results for these two tests.

To test the efficacy of this system to work in real time and on different users, four people, none of whom contributed training data for the models, performed 15 sequences of the five *T'ai Chi* gestures[1]. The tracking data from *STIVE* was smoothed and resampled in real-time[2], and the data was fed into a real-time Viterbi algorithm[3]. Table 5.2 shows the confusion matrix of the recognition results.

5.2 Identifying Salient Points in the Gesture

The experiment to test the efficiency of the system at identifying the salient moments in the users' gestures is slightly less direct. As was discussed in the previous chapter, it is difficult to assign a precise definition for what are such moments in the gesture. The Viterbi algorithm returns the moments in the gesture where the model for the given gesture had the highest and lowest probability of producing the given observation sequence. How

[1] Why only five *T'ai Chi* gestures? In our previous work [11], we developed models for eighteen different gestures. However, because of the different preprocessing involved in this real-time implementation, we could not simply use the same models without completely retraining. Because we wanted *Sensei* to demonstrate the ability to recognize gestures in a natural sequence, the five gestures we chose are the first five gestures in the short, Yang sequence [41]. Included in this sequence were two gestures, "Opening Form" and "Grab the Bird's Tail" which were commonly confused in our previous system.

[2] See Appendix B for details about the smoothing and resampling.

[3] See Appendix A for further notes on the implementation of the Viterbi that was used.

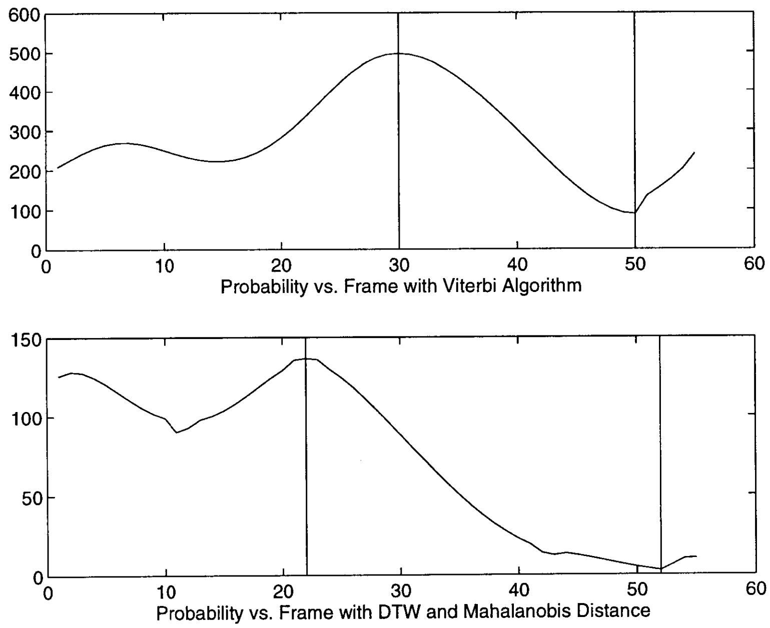
do these moments relate to the moments given by the definition – the moments where the movement of the user's hands most differed from the movement of the hands in the ideal version of the gesture? If ground truth data on the location of those moments was known, we could simply compare the moments that gave the minimum and maximum probabilities in the Viterbi calculation to this ground truth data.

Although that ground truth data does not exist, it is possible to arrive at an acceptable approximation. Consider again the simple definition for the salient moments as described in the previous chapter: the salient moments are the instances in the user's gesture where the hands were furthest in space from where the hands were in the ideal version of the gesture, with some flexibility in correspondences. We discarded this definition because it didn't allow for shift invariance or invariance to the size of the users' bodies. However, what if the user *is* the expert, and the user *does* sit in the same position as she did when she trained the system? Then it would be possible, using a simple dynamic time warping algorithm which uses Mahalanobis distance as a distance metric, to locate the moment in the user's gesture when the hands were most different and least different from the ideal positions. If we use these moments as our ground truth data, they can then be compared to the values determined with the Viterbi algorithm running in our normal, velocity space.

To clarify this procedure, consider Figure 5-1. In this figure, the probability track for an actual "Opening Form" gesture as computed with the Viterbi algorithm is shown on top. Vertical lines show the locations of the moments classified as "best" and "worst." The lower part of the figure shows the same gesture as segmented with a dynamic time warping algorithm using Mahalanobis distance as a distance metric. Figure 5-2 shows how we use the DTW algorithm as ground truth to compute an error percentage for the classification from the Viterbi algorithm.

To evaluate the performance of this aspect of the *T'ai Chi Teacher*, we devised two experiments. The goal of the first was to test the system's ability to find the types of errors typical of beginners: gross movements obviously out of the correct trajectory. The goal of the second experiment was to test the system's ability to find the more subtle types of errors typical of experts.

The HMM's for both experiments were trained on data from a single user. Thirty examples of each gesture were performed by the user sitting in the same position. For each of the two experiments, ten different examples of each gesture were performed by the same



Probability vs. Frame with Viterbi Algorithm

Probability vs. Frame with DTW and Mahalanobis Distance

Figure 5-1: The top view is the probability/frame along the observation sequence computed by the Viterbi algorithm. The lower view is the same sequence, but with probability computed by a DTW algorithm using Mahalanobis distance as a distance metric. The gesture was "Opening Form." At approximately frame 52, the user moved his left hand well out of the expected trajectory.

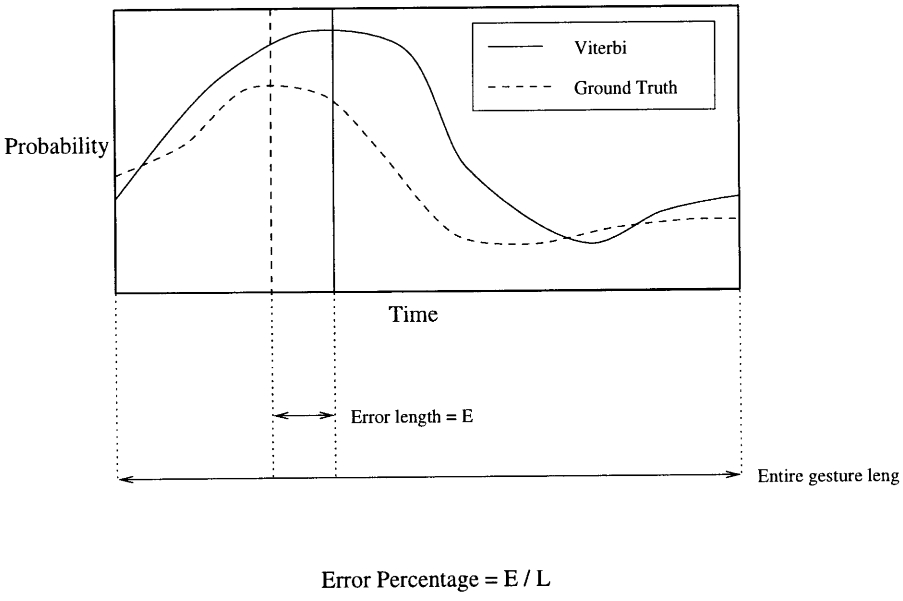


Figure 5-2: Hypothetical probability tracks as computed with the two methods. The vertical lines show the classification as calculated in both methods. We assume the classification from the DTW algorithm is ground truth and compute an error on the classification of the Viterbi algorithm. The error percentage is the time difference of the two classifications divided by the time of the entire gesture.

experiment	*Bad moments*	*Good moments*
beginner	5%	12%
expert	15%	14%

Table 5.3: Errors in classification of the instants in the gesture where the user most (good moments) or least (bad moments) matched the ideal gesture. Ground truth is provided by the classification from the DTW algorithm. The error percentages are calculated by dividing the time difference of the classification from the Viterbi algorithm and the ground truth classification by the time length of the entire gesture. The beginner experiment included gestures where the user made large scale, beginner-type errors, and the second experiment was with expert-type errors.

user, sitting in the same location. The moments chosen with the Viterbi algorithm were compared to those generated by a dynamic time warping system which used Mahalanobis distance as a distance metric. Table 5.2 shows the average time between the two instances chosen by both methods over the course of the testing data for both experiments.

5.3 Analysis of Results

When considering the recognition results, some important *caveats* are in order. In the first two experiments, recognition was tested on sequences which had been presegmented into sentences of five gestures. In this case, the classification of an observation sequence is simply whichever model has the highest probability of having emitted that sequence. In the real-time, multiple user experiment, a different definition of classification is needed because the data is not segmented. The procedure used is as follows:

- As described in the previous chapter, run the Viterbi algorithm over a window of frames as each new frame is observed.
- Take the maximum $Pr(\lambda)$ over all the models and this is the most likely model at that frame.
- If the probability for that model exceeds its set threshold, classify the sequence as an instance of the model.

When setting the thresholds, there is a tradeoff between insertion and deletion errors. The thresholds used in *Sensei* were set before the experiments by testing models trained on all 60 examples in the training data on independent example data and minimizing the sum

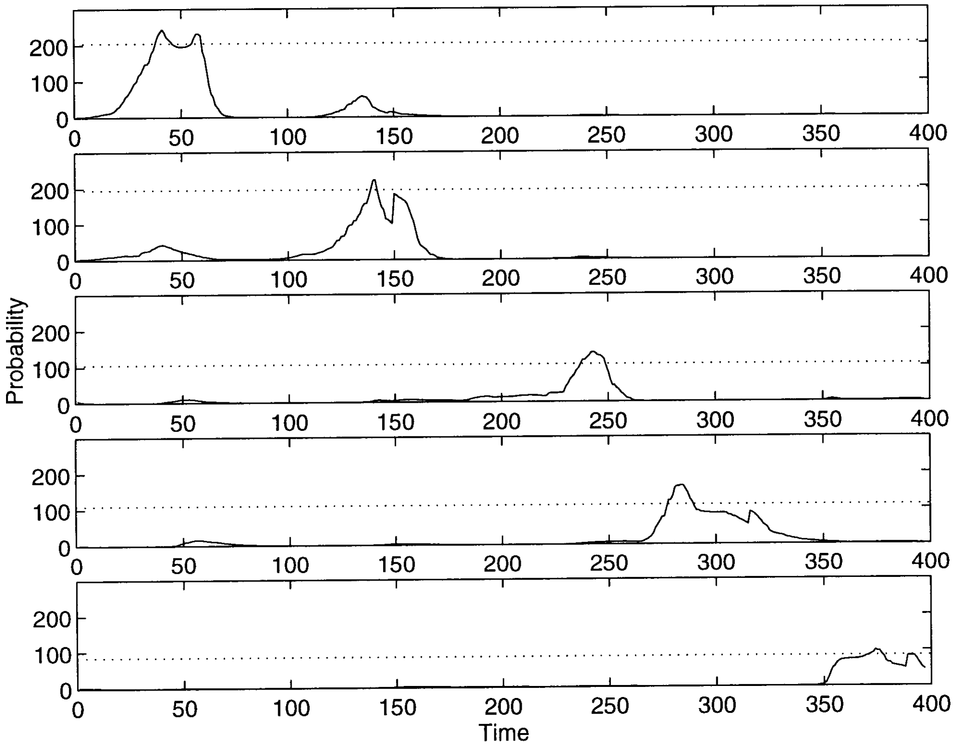


Figure 5-3: The gestures for each plot are, from top to bottom, Opening Form, Grab the Bird's Tail, Wave Hands Like Clouds, Single Whip, and Brush Knee. The dotted lines are the preset thresholds. Classification occurs when the probability for one of the gestures eclipses its threshold.

of the insertion and deletion errors. Figure 5-3 shows an example sequence, the thresholds, and the derived classifications.

A recurrent problem originally faced in these experiments was that the probabilities for all the models rose when the user paused, holding the hands still. The reason this happens is that all of the models have states with means near the origin. It is possible, with its dynamic time warping, for the Viterbi algorithm to explain an observation sequence by assuming that a model was in this state with low velocity for most of the observation sequence. The probability per frame, in that case, will rise. In order to alleviate this problem, we included a silence model: an HMM that modeled non-moving hands. The testing data used in the experiments did not typically contain pauses, so the silence model has been omitted from the results. In practice, however, the addition of this model into the recognition scheme was found to be effective.

Another problem in the classification scheme discussed above is that when looking for the maximum $Pr(\lambda)$ over all the models, we assumed that the probabilities were uniform for all the gestures. In reality, they are not. The different models have different numbers of states, each of which has a unique probability density function with its own scale. One method to handle this problem is to use cross-validation in the training stage. By adjusting the parameters of the models, one can minimize the differences between models in the probabilities of the training data.

Finally, it is clear that the Viterbi algorithm is not perfect in analyzing the motions of the gestures to find the salient moments. The system is effective at locating the gross, large-scale types of errors typical of beginners, but not as effective at locating the more subtle errors typical of experts, as the results in 5.2 shows.

One problem is simply in finding the right definition for what the salient moments are. One piece of future work is to have human experts observe a user and identify these moments by hand. This classification can then be compared to that done by the Viterbi algorithm in an effort to learn the "correct" classifications.

Another drawback to the Viterbi comes simply because HMM's approximate the observation space with a small number of states. For example, a model with three states essentially forces the observation sequence into the three probability density function blobs along the trajectory. The probability dips as the observations move from one state to the next, until the observation more closely matches the mean of the new state. More states in the model, of course, will help to alleviate this problem. This situation can be most easily seen in the results for the gesture "wave hands like clouds." This gesture is performed with circular movements of the hands. When there is a continual change in the direction of motion, as in circular movements, models in velocity space need a large number of states or this quantizing error will increase. We found the errors in the experiments for "wave hands like clouds" to be almost twice as large as those of the other gestures.

Chapter 6

Summary and Future Work

6.1 Summary

The real-time, interactive feedback system, *Sensei: The T'ai Chi Teacher*, has been shown. This application provides a good platform on which to build a more sophisticated teaching, training, and feedback tool for gestures or action. To support this thesis, we define a good platform to be one in which successful gesture recognition is achieved and in which salient moments in the gesture are identified. With these two tools, a more sophisticated teacher can understand what gesture the user is attempting to perform and at what moments in the action the user should receive feedback.

The gesture recognition task is accomplished through the use of Hidden Markov Models. A recognition rate of 95.2% was achieved on multiple users performing five gestures in real-time. The same HMM framework is used to locate the salient moments in a gesture. An average error rate of 8% for larger, beginner moves and 14% for more subtle, expert moves was achieved in locating these moments, when compared to the instants calculated by comparing the motion trajectory of the hands to that of an expert.

6.2 Future Work

What the *T'ai Chi Teacher* does not purport to be is a complete *T'ai Chi* gesture recognition system. Currently, it is only capable of recognizing five different gestures. One simple extension of this work would be to increase the number of gestures in the system's vocabulary. The success of our previous *T'ai Chi* recognition system [11] suggests this goal is easily attainable. In addition, using the *STIVE* tracking system does not allow the system to distinguish between fine differences in hand positions and orientations which are an important part of *T'ai Chi* movements. As the tracking system improves, features based on the orientation of the hands might prove to be a valuable addition to the models. Alternatively, a foveated camera driven by the hand tracking would also accomplish this goal.

Another extension would be to grow the system into a general action feedback and teaching tool. The idea here is to turn the application into one which allows experts in a variety of gestures or actions, such as American Sign Language or tennis strokes, to train the system. The same framework demonstrated here would extend naturally to other vocabularies; by generalizing the procedure of training the HMM's, this application could be a general action teacher, rather than a *T'ai Chi* Teacher.

The feedback elements of the system could also be extended in several directions. A system could be developed that models the ability level of the user and adjusts the type of feedback that it gives accordingly. For example, a beginner in the art of *T'ai Chi* would probably be better served by receiving an abundance of positive feedback until she becomes more proficient. An expert, on the other hand, would probably best benefit from a highly critical feedback, one that identifies fine-grained differences between the motion and the ideal gesture.

Also, the growing bed of tools provided by the field of affective computing [23] could be utilized to tune the feedback provided by the teacher. For example, if a user starts to become frustrated with the system, the frequency of interruptions and feedback could be adjusted.

Finally, one piece of further work for the feedback that could make the teaching particularly useful would be to use the display screen to not only show the user where her hands are at any given time but where they "should" be at that moment. This allows the user to modulate her motion as she is performing the gesture, as opposed to receiving feedback only upon its completion.

Appendix A

Computational Advantage of Training the Models Backwards

In Chapter 4 we described the process of running the Viterbi algorithm over a range of sequence lengths. One disadvantage of this process is that in the standard implementation of the Viterbi algorithm, the calculation of the entire lattice of probabilities depends on the values of the probabilities in the first frame. Recall that the procedure is essentially recursive, with all subsequent probability calculations depending on the previous ones. To run the Viterbi over sequences of varying lengths, where the first frame changes from T_{max} to T_{min} while the last frame is held fixed, necessitates recomputing the Viterbi lattice for each sequence length. To avoid this computationally expensive procedure, we time reverse the data. Each new frame becomes, in effect, the first frame in the sequence. Then, when the Viterbi algorithm is run over sequences of varying lengths, the first frame is stationary, but the last frame moves from T_{max} to T_{min}. The lattice is simply computed once for the maximum length and it then contains all the probabilities needed for the entire sequence.

This modification, however, implies that the models must be trained on data that is similarly time reversed. Training data is recorded and then flipped in time before the models are trained. The resulting beginning state, then, is in actuality the final state, and vice-versa. In addition, the prior probability, π, becomes the closing probability (the probability of the final state being state n)[1].

[1] The idea presented in this appendix was originally suggested by Andrew Wilson.

Appendix B

Smoothing the Data

B.1 Filtering

The hand and head tracking data provided by *STIVE* is subject to noise. The velocity features we use to do recognition are by nature particularly susceptible to noise, so it is important to low-pass filter the data to alleviate the problem. Because the movements involved in *T'ai Chi* gestures are slow and smooth, there is little risk of losing salient information by filtering. The filter we use is a 23 tap minimax filter with a cut-off frequency of 3Hz. This introduces a 12 tap lag (which corresponds to 0.36 seconds at our sampling rate of 30Hz.) in the processing. However the benefit in the calculation of the velocity compensates for this negative attribute. Figure B-1 shows a plot of the raw data and the smoothed data.

B.2 Resampling

Hidden Markov Models assume that the observations are evenly sampled in time. This is not the case with *STIVE* data. The *T'ai Chi Teacher* takes data incoming from the *STIVE* system and resamples it in time at a rate of 30Hz. The goal is this: given a set of function values $f(x_i)$ at locations x_i for $i = 1 \ldots N$, determine the value of $f(x)$. Several methods exist for this task, differentiated by the constraints placed on the interpolated function. We use a natural cubic spline interpolator: one in which the first derivative is smooth and the second derivative is continuous. Furthermore, the second derivative is constrained to be 0 at the boundaries ($i = 1, N$).

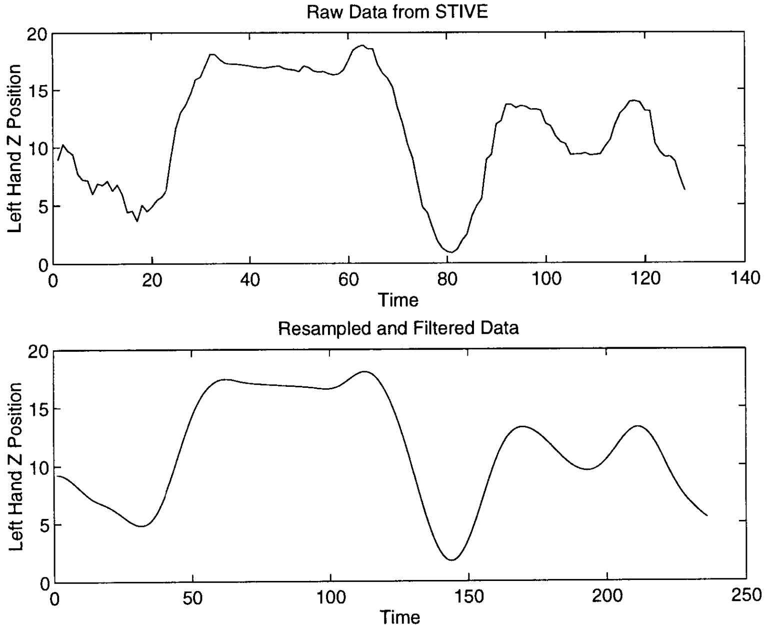


Figure B-1: An example of the raw and resampled/filtered data.

Because we already cache 23 samples of data for the filter, we use this same segment to compute resampling ($i = 1 \ldots 23$). The cubic spline algorithm we use is a standard implementation from [25].

Bibliography

[1] Jeanne Achterberg and G. Frank Lawlis. *Imagery of Disease.* Institute for Personality and Ability Testing, Champaign, IL, 1978.

[2] H. V. Allington. Review of psychotherapy of warts. *Arch Dermatol Syphil*, 66:316–326, 1952.

[3] R. Asher. Respectable hypnosis. *British Medical Journal*, 2:309–313, 1956.

[4] Ali Azarbayejani and Alex Pentland. Real-time self-calibrating stereo person tracking using 3-D shape estimation from blob features. In *Proceedings of 13th ICPR*, Vienna, Austria, August 1996. IEEE Computer Society Press.

[5] Lee Baer and Owen S. Surman. Microcomputer-assisted relaxation. *Perceptual and Motor Skills*, 61:499–502, 1985.

[6] L. Baider, B. Uziely, and A. K. De-Nour. Progressive muscle relaxation and guided imagery in cancer patients. *Gen. Hosp. Psychiatry*, pages 340–347, Sept 1994.

[7] David A. Becker and Alex Pentland. Staying Alive: A virtual reality visualization tool for cancer patients. In *Proceedings of the AAAI'96 Workshop on Entertainment and Alife/AI*, 1996.

[8] Herbert Benson. *The Relaxation Response.* William Morrow and Co., New York, NY, 1975.

[9] B. Block. About the curing of warts by suggestion. *Klin Worchenschr*, 6:2271–2325, 1927.

[10] L. R. Bridge, P. Benson, P. C. Pietroni, and R. G. Priest. Relaxation and imagery in the treatment of breast cancer. *British Medical Journal*, pages 1169–1172, Nov 1988.

[11] L. W. Campbell, D. A. Becker, A. J. Azarbayejani, A. F. Bobick, and A. Pentland. Invariant features for 3-D gesture recognition. In *Second International Conference on Face and Gesture Recognition*, pages 157–162, Killington, VT, Oct 1996. IEEE Computer Society Press.

[12] L. Couper and T. Davies. A difficult wart treated by suggestion. *British Medical Journal*, 2:1398, 1952.

[13] T.J. Darrell and A.P. Pentland. Space-time gestures. *Proc. Comp. Vis. and Pattern Rec.*, pages 335–340, 1993.

[14] T. W. Decker, J. Cline-Eisen, and M. Gallagher. Relaxation therapy as an adjunct in radiation oncology. *Journal of Clinical Psychology*, pages 388–393, May 1992.

[15] C. S. Feldman and H. C. Salzberg. The role of imagery in the hypnotic treatment of adverse reactions to cancer therapy. *J. S. C. Med. Assoc.*, pages 303–306, 1990.

[16] Y. He and A. Kundu. Planar shape classification using Hidden Markov Models. In *Proc. IEEE Conf. on Comp. Vision and Pat. Rec.*, pages 10–15. IEEE Press, 1991.

[17] Caryle Hirshberg and Marc Ian Barasch. *Remarkable Recovery.* Riverhead Books, New York, NY, 1995.

[18] M. H. Hockenberry. Guided imagery as a coping measure for children with cancer. *Journal of Assoc. Pediatr. Oncol. Nurses*, page 29, 1989.

[19] X.D. Huang, Y. Ariki, and M. A. Jack. *Hidden Markov Models for Speech Recognition.* Edinburgh University Press, 1990.

[20] Bill Little. *Eight Ways to Take an Active Role in Your Health.* Harold Shaw Publishers, Wheaton, IL, 1995.

[21] M. McDowell. Juvenile warts removed with the use of hypnotic suggestion. *Bull. Menninger Clin.*, 13:124–126, 1949.

[22] A. M. Memmesheimer and B. Eisenlohr. Surveys about the treatment by suggestion of warts. *Dermatol*, 62:63–68, 1931.

[23] R. W. Picard. Affective computing. MIT Media Lab Perceptual Computing Group Technical Rep ort No. 321, Massachusetts Institute of Technology, 1995.

[24] J. Post-White. The effects of imagery on emotions, immune function, and cancer outcome. *Mainlines*, 14(1):18–20, Winter 1993.

[25] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C.* Cambridge University Press, 1992.

[26] L. R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc. of the IEEE*, 77(2):257–285, February 1989.

[27] L. R. Rabiner and B. H. Juang. An introduction to Hidden Markov Models. *IEEE ASSP Magazine*, pages 4–16, Jan 1986.

[28] D. A. Rapkin, M. Straubing, and J. C. Holroyd. Guided imagery, hypnosis and recovery from head and neck cancer surgery: an exploratory study. *International Journal of Clinical Experimental Hypnosis*, pages 215–226, Oct 1997.

[29] J. Schlenzig, E. Hunter, and R. Jain. Recursive identification of gesture inputs using Hidden Markov Models. *Proc. Second Annual Conference on Applications of Computer Vision*, pages 187–194, Dec 1994.

[30] Carl O. Simonton, Stephanie Matthews-Simonton, and James L. Creighton. *Getting Well Again.* J. P. Tarcher, Los Angeles, CA, 1978.

[31] S. E. Sims. Relaxation training as a technique for helping patients cope with the experience of cancer: a selective review of the literature. *Journal of Advanced Nursing*, pages 583–591, Sept 1987.

[32] Nicholas P. Spanos, Robert J. Stenstrom, and Joseph C. Johnston. Hypnosis, placebo, and suggestion in the treatment of warts. *Pyschosomatic Medicine*, 50(3):245–260, 1988.

[33] Nicholas P. Spanos, Victoria Williams, and Maxwell I. Gwynn. Effects of hypnotic, placebo, and salicylic acid treatments on wart regression. *Pyschosomatic Medicine*, 52(1):109–114, 1990.

[34] David Spiegel, Joan R. Bloom, Helena C. Kraemer, and Ellen Gottheil. Effect of psychosocial treatment on survival of patients with metastatic breast cancer. *The Lancet*, pages 888–891, October 1989.

[35] T. E. Starner and A. Pentland. Visual recognition of American Sign Language using Hidden Markov Models. In *Proc. of the Intl. Workshop on Automatic Face- and Gesture- Recognition*, Zurich, 1995.

[36] A. Stolcke and S. Omohundro. Hidden Markov Model induction by Bayesian model merging. In *Advances in Neural Information Processing Systems*, pages 11–18. Morgan Kaufman, 1992.

[37] O. S. Surman, S. K. Gottlieb, T. P. Hackett, and E. L. Silverberg. Hypnosis in the treatment of warts. *Archives of General Psychiatry*, 28:439–441, March 1973.

[38] M. Ullman and S. Z. Dudek. On the psyche and warts. *Pyschosomatic Medicine*, 22:68–76, 1960.

[39] A. D. Wilson and A. F. Bobick. Learning visual behavior for gesture analysis. In *Proc. IEEE Int'l. Symp. on Comp. Vis.*, Coral Gables, FL, November 1995.

[40] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using Hidden Markov Model. *Proc. Comp. Vis. and Pattern Rec*, pages 379–385, 1992.

[41] Shing Yen-Ling. *T'ai Chi Ch'uan: The Basic Exercises.* Sugawara Martial Arts, 1990.