Visual Classification of Co-Verbal Gestures for Gesture Understanding

by

Lee Winston Campbell

B. A. in Physics, Middlebury College (1978)
M. S. in Media Arts and Sciences, Massachusetts Institute of Technology (1994)

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

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2001

Author .....
Program in Media Arts and Sciences,
School of Architecture and Planning,
May 4, 2001

Certified by .....
Aaron F. Bobick
Associate Professor of Computational Vision
Georgia Institute of Technology College of Computing
Thesis Supervisor

Certified by .....
Justine Cassell
AT&T Career Development
Associate Professor of Media Arts and Sciences
Thesis Supervisor

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

Visual Classification of Co-Verbal Gestures for Gesture Understanding

by

Lee Winston Campbell

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
on May 10, 2001, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

A person's communicative intent can be better understood by either a human or a machine if the person's gestures are understood. This thesis project demonstrates an expansion of both the range of co-verbal gestures a machine can identify, and the range of communicative intents the machine can infer. We develop an automatic system that uses realtime video as sensory input and then segments, classifies, and responds to co-verbal gestures made by users in realtime as they converse with a synthetic character known as REA, which is being developed in parallel by Justine Cassell and her students at the MIT Media Lab.

A set of 670 natural gestures, videotaped and visually tracked in the course of conversational interviews and then hand segmented and annotated according to a widely used gesture classification scheme, is used in an offline training process that trains Hidden Markov Model classifiers. A number of feature sets are extracted and tested in the offline training process, and the best performer is employed in an online HMM segmenter and classifier that requires no encumbering attachments to the user. Modifications made to the REA system enable REA to respond to the user's beat and deictic gestures as well as turntaking requests the user may convey in gesture.

The recognition results obtained are far above chance, but too low for use in a production recognition system. The results provide a measure of validity for the gesture categories chosen, and they provide positive evidence for an appealing but difficult to prove proposition: to the extent that a machine can recognize and use these categories of gestures to infer information not present in the words spoken, there is exploitable complementary information in the gesture stream.

Thesis Supervisor: Aaron F. Bobick
Title: Associate Professor of Computational Vision
Georgia Institute of Technology College of Computing

Thesis Supervisor: Justine Cassell
Title: AT&T Career Development
Associate Professor of Media Arts and Sciences

Doctoral Dissertation Committee

Thesis Supervisor.....

Aaron F. Bobick

Associate Professor of Computational Vision

Georgia Institute of Technology College of Computing

Thesis Supervisor.....

Justine Cassell

Associate Professor of Media Arts and Sciences

Program in Media Arts and Sciences

Thesis Reader .....

Thomas S. Huang

Professor of Electrical and Computer Engineering,

Beckman Institute, University of Illinois at Urbana-Champaign

Acknowledgments

There are many people to whom I am deeply grateful for the opportunities I've been lucky enough to have here at the MIT Media Lab. First, I thank my advisors. Professor Aaron Bobick and Professor Justine Cassell. They have supported me, advised me, set deadlines for me, and in general coaxed better work from me than I ever thought I was capable of. I also thank my outside thesis reader Professor Thomas Huang who has been generous with his time and wisdom. The faults of this thesis are mine; the good work in it could not have been done without my advisors and committee.

Next, I thank my fellow students in both the GNL and Vismod groups. Much of my grad school learning happened student to student, and I have been lucky enough to be placed with a marvelous group of people who generously shared their time and wisdom. My favorite project while at the Media Lab was the Kidsroom, done by Aaron's HLV group, where I learned a great deal. I thank my fellow HLVers and collaborators Stephen Intille, Claudio Pinhanez, Yuri Ivanov, Andy Wilson, and Jim Davis, as well as Freedom Baird and the outside collaborators who all helped make the project such a wonderful group experience. I also thank Chris Wren, Ali Azarbayejani, Thad Starner, Dave Becker, Hannes Vilhjalmsson, Tim Bickmore, Hao Yan, Kimiko Ryokai, Kenny Chang, Jennifer Smith, Petra Chong, Mark Billinghurst, Yukiko Nakano, Erin Panttaja, and Sola Grantham for great learning experiences and collaborations on REA and other projects.

I learned a great deal from Vismodders who were senior when I started, including Irfan Essa, Nassir Navab, Kris Popat, Fang Liu, Sourabh Niyogi, Trevor Darrell, Stan Sclaroff, Marty Friedmann, Bill Freeman, John Wang, Alex Sherstinski, Bradley Horowitz, Monika Gorkani, Ali Rahimi, Pawan Sinha, Dan Ellis, Sandy Pentland, Ted Adelson and Roz Picard. I learned even more working with my contemporaries, many mentioned above, but I also thank Tom Minka, Sumit Basu, Tony Jebara, Ken Russell, Chahab Nastar, Baback Moghaddam, Flavia Sparacino, Martin Szummer, Matt Krom, Elizabeth Sylvan, Bradley Rhodes, Lenny Foner, Raul Fernandez, Brian Clarkson, Deb Roy, Nuria Oliver, Jocelyn Scheirer, Jennifer Healey. I've found staff and affiliates very kind and helpful, and I thank Hisashi Aoki, Erik Trimble, Dave Berger, David Long, Andrew Donnelly, Judy Bornstein, Karen Navarro, Kate Mongiat, Robin Simone, Glen Sherman, Viet Anh, Jane Wojcik, and Chi Yung Yuen. I owe a special debt of gratitude to the folks at the Leg Lab and elsewhere who inspired and encouraged me to apply to graduate school, including Robert Lees, Marc Raibert, Frank Vallese, Kevin Grimm, Robert Ringrose, Rob Playter, Charles Francois, Ed McCluney, and Carolyn Akinbami.

Finally, I am eternally grateful to my parents, Malcolm and Jeanne Campbell, and my wonderful wife and partner in life Ceci Dunn, and our darling son Benjamin Dunn Campbell. My parents always encouraged my curiosity and supported my endeavors and now as a parent myself, I find the only possible way to thank them is to try to raise our children as well as they did us. My wife has been my island of sanity and my constant support throughout this thesis to the extent that I doubt I would have finished it without her. Our 18 month old son Benjamin is a wellspring of love and joy, and only by looking at the world through his eyes was I able to say with conviction "it's only a thesis!"

Contents

List of Figures

Chapter 1

Introduction

People make hand gestures as they talk. Some gestures, such as "V for victory" or "thumbs up" have a particular hand shape or motion associated with them. Other gestures have no right or wrong way to be performed. For example, people often move their hands to emphasize a word or syllable. Or they may motion between two points in space while stating an "either or" contrast. We define such gestures as co-verbal gestures.

Discourse, as used in this work, is defined as the interactive communication of one person with another or with a machine. The study of discourse is distinct from linguistics because it addresses the meaning of whole conversations and of the nonverbal as well as verbal components of the conversation, and is concerned with non-grammatical and disfluent speech, in addition to well formed sentences.

1.1 Statement of the problem

In this thesis project I have developed an automatic system to segment, classify, and respond to co-verbal gestures made by users in realtime in the course of conversation with a synthetic character known as REA, which uses a linguistic grammar approach to both generation and understanding of multimodal conversation. The challenges are twofold: first, to determine what kinds of visual features are characteristic of the semantics of gestures; and second, to exploit the gesture information to aid in understanding conversation.

The purpose of labeling gestures is to use them as an additional information source to supplement speech. For this effort to be feasible, two questions must be answered: (a) can gesture information supplement speech? and (b) is classification feasible? This thesis demonstrates that the answers to both questions is a qualified "yes," and articulates a method of of achieving these goals in the context of REA's grammar based understanding system.

In addition, the methodology used by gesture researchers has not strongly established that gestures provide an information stream that augments speech. Gesture classification has always been done by humans while listening to the associated speech. Although the gesture stream is studied for its contribution to the understanding of speech, this methodology leaves open the opposite interpretation: that for many gestures, speech is primary, and contributes to the understanding of gesture. The methodology of this thesis – building a system that classifies without reference to speech, then inferring communicative intent from gestures and their co-occurrence with speech, and then altering REA's behavior to better respond to the communicative intent – provides objective evidence that gestures carry information that supplements the associated speech.

1.2 Why are gestures important?

Gestures are a component of the most natural form of communication known to humans – face to face conversation. Gestures accompany over 75% of all clauses in face to face conversation [35]. Understanding gestures and identifying their co-occurrence with speech are capabilities important to builders of automated systems who strive to achieve that same degree of efficiency and naturalness in communication. Complementary gestures carry information not present in speech; so an interface that understands gestures as well as speech will be more effective than a speech-only interface. The importance of non-verbal communication becomes readily apparent in telephone conference calls. In two-person telephone conversations turn taking signals can transmitted verbally, but any one who has participated in a conference call has experienced the many false starts, interruptions, and long pauses that result when non-verbal turntaking channels are not present. Rogers [46] found gestures were relied on to assist speech understanding in noisy settings.

Gestures, their timing, and the context in which they occur can convey information about the topic under discussion; for example: pointing to an object and saying "I want this." Thus a system that can recover some of this gestural information will be a better listener than a system that ignores gesture. Gestures also carry information about the protocols of conversation; for example: listeners often bring their hands up into a "prepared to gesture" position to signal a desire to speak. Thus a system that attends to these protocols will be more efficient at responding and more pleasant to talk to than a system which is unaware of the protocols.

1.3 Outline of the thesis

This project is an effort to classify gestures in order to use them as an independent information source to supplement speech. For this effort to be feasible, two questions must be answered: (a) can gesture information supplement speech? and (b) is classification feasible? These questions are addressed in section 6.9 and 6 respectively.

Section 2 shows why this problem is of interest to the community of discourse researchers, and lays out the groundwork of discourse for readers less familiar with the field.

Section 3 places this work in the context of related work in discourse and in understanding human body movement.

Section 4 describes the classification methodology used in this project, and the reasons behind the choices made.

Section 5 describes how the gesture recognition system was trained on naturally occurring gestures, and how the realtime gesture classification system was implemented and integrated into REA.

Section 6 addresses question (b) above, presenting the results of training and testing the gesture recognition system.

Section 6.9 describes some applications of recognition / classification, and addresses question (a) by showing how behaviors like emphasis gestures, pointing gestures, and turntaking cues can be used to add information to that available from speech.

Section 7 concludes.

Chapter 2

Motivation

There are two main motivations for this thesis: one is an application – to build an essential component of a multi-modal conversational character of the sort envisioned by Cassell and her collaborators [12]. The other is theoretical – to delve into what information is transmitted by gestures.

The Media Lab's Gesture and Narrative Language group has been working since 1998 on a project called REA, one of the goals of which is to study and test theories about the protocols of conversation. Gestures are an essential part of these protocols, and this thesis builds a vision based gesture classification system that generates high level gesture descriptions as input to the rest of the REA project.

Co-Verbal gesture studies since the 1940s have developed two main approaches for classifying co-verbal gestures: by their inherent characteristics, i.e. morphology; and by their function in the conversation - the communicative intent they transmit. An interesting scientific question is: are there morphologic features that correspond to particular functions – does morphology indicate function and communicative intent? This thesis work will be the first project to address these questions by making a computational model of gesture classification.

2.1 Background on co-verbal gestures

The fact is that people have natural conversations on the telephone every day; therefore, gestures are not a necessary component of interactive communication. Conversely, many people make hand gestures even when speaking on the phone. So what functions do hand gesture serve? What information is carried in them that is not present in speech? When the gesture channel is available, many people use it. In fact, some kinds of conversations, for example asking driving directions, are significantly more difficult without hand gestures. As will be explained, gestures convey information about the topic of conversation, and about turntaking and similar behaviors.

2.1.1 Some Terminology

Before going further, it is helpful to give some rough definitions and quick examples of what is meant by co-verbal gestures – the gestures associated with speech. Discourse researchers have converged on four main categories of co-verbal gestures, and following McNeill [35] I've included Butterworths as a fifth. After McNeill, we will use these terms:

- *Beats* – brief hand motions (often downward) in which hand shape is not controlled; they often convey emphasis.
- *Deictics* – pointing gestures, which may refer either to a physical object; e.g. "that window" or a concept, e.g. "the operating system."
- *Iconics* – miming of actions or drawing shapes of concrete things; e.g. saying "the lawn is as flat as a billiard table," while smoothing the surface of an imaginary table.
- *Metaphorics* – just like iconics but referring to abstract things; e.g. saying "a very hierarchical organization," while drawing a pyramid in the air
- *Butterworths* – repeated, shaking gestures often made while searching for a word; often accompanied by filled pauses (e.g. "umm..., ahh...").

In addition, there are several categories of gestures which are independent of speech and thus not classified as co-verbal gestures:

- *Adjustors* – gestures such as rubbing one's face or scratching an itch.
- *Emblems* – gestures such as 'Aok' or the rich families of obscene gestures which have culturally assigned meanings.
- *Sign languages* – languages evolved for communication among the deaf, with large vocabularies, complex grammars, and a notion of correct formation for each word.

Not only unconnected with speech, these last three categories of movements are distinct from co-verbal gestures in other ways. First, adjustors are usually performed with no communicative intent. Of course one can pantomime scratching an embarrassing area for humorous purposes, or arrange to make e.g. nose rubbing a signal, but that is not the native form of adjusters.

Emblems and sign languages, on the other hand, are highly communicative, but they are essentially linguistic: well formed gestures that can be used instead of words. They are distinct and atomic and usually context insensitive, and can be combined into larger grammatical units. Significant progress has been made in automatic recognition of sign languages; see [53, 51] for recognition from visual input, and [33] for dataglove input. In contrast, co-verbal gestures, in McNeill's terminology, "have no standards of form;" are context sensitive and non-combinatoric; and are "global and synthetic," which means that the parts of a gesture are determined by the whole gesture, unlike e.g. sign language sentences where the parts (words) determine the meaning of the whole. These non-linguistic properties make co-verbal gestures a distinct field of study, with different methods than the study of linguistic gestures such as emblems and sign languages.

Another way of dividing gestures is by the discourse functions they serve. The major functions are:

- *Feedback* – signaling "I understand, keep going" or "I'm confused; elaborate more," usually by head nods or shakes, or "mmhmm" sounds.
- *Turntaking* – can be subdivided into turn requesting, holding, and yielding; requesting is often signaled by beginning to gesture; holding by continuing to gesture or make

"umm, uhh" sounds or both; and yielding by ending both speech and gesture. Glances also play a role in turntaking.

- *Information Structure* – components called 'theme' and 'rheme:' themes are topics already part of the shared context; rhemes are new or spotlighted. For example: in "I saw the boy with a dog," 'the' signals theme – the boy is already known; while 'a' indicates rheme – the dog is new to the discourse. Emphasis, either intonational or via a beat gesture, is another rheme indicator. Section 6.11.2 contains more examples of themes and rhemes.

- *Contrast* – either / or choices, for example, "tile or wood" are often expressed with a dual emphasis.

As described in the examples, these discourse functions are served by multiple modalities; most often gesture, intonation, facial expression, and head movement. In the current state of the art, intonation is very hard to recover; even the problem of classifying sentences into statements and questions is still an open research problem [36]. So gesture classification can make a significant contribution to the state of the art of computer understanding of discourse functions.

This functional view of gestures is the most important for the goals of this thesis. The REA project is an effort to map multiple modalities such as speech, gesture, and intonation onto discourse functions, both for input to REA, and for output generated by REA. Thus identifying the discourse functions of gestures provides a natural and principled way to combine gestures with the other modalities.

2.2 When gestures are not redundant to speech

According to Ekman & Friesen [21], the meaning of a gesture can be informative, communicative, or interactive. We denote the first two groups as content, and the last as interactive. Content and interactive gestures serve different discourse functions, so distinguishing between them will be an important part of this thesis. The next sections will show how the meanings of gestures can be divided along this axis.

Figure 2-1: Subject DS saying "you could just have a beautiful view of the lake" while spreading her hands to the sides during "beautiful view." This two handed iconic gesrtures indicates the breadth of the view, conveying information not present in the speech. In this video image, the two views in the upper quadrants are from the two cameras of the tracking system. The three blobs in the bottom left quadrant show the output of the tracking system. The bottom right quadrant is blank.

2.2.1 Content gestures

Sometimes gestures carry topic information not present in the speech. Ekman [21] called these illustrators; we will denote them as content gestures. For example, when one says "put that there" while pointing first at an object, then at a place, the meanings of both "that" and "there" are found in the deictic (pointing) gestures, not the speech. In 1984 Bolt [8] developed a system that would understand and obey "put that there" sentences in a specific domain. [1] In figure 2.2.1 subject DS is describing a beautiful view. Her two handed iconic gesture conveys information about the breadth of the view not present in the dialog.

[1]Essentially, it used a placeholder word in the speech, such as "this" or "that" or "there", to signal when to attend to gestures.

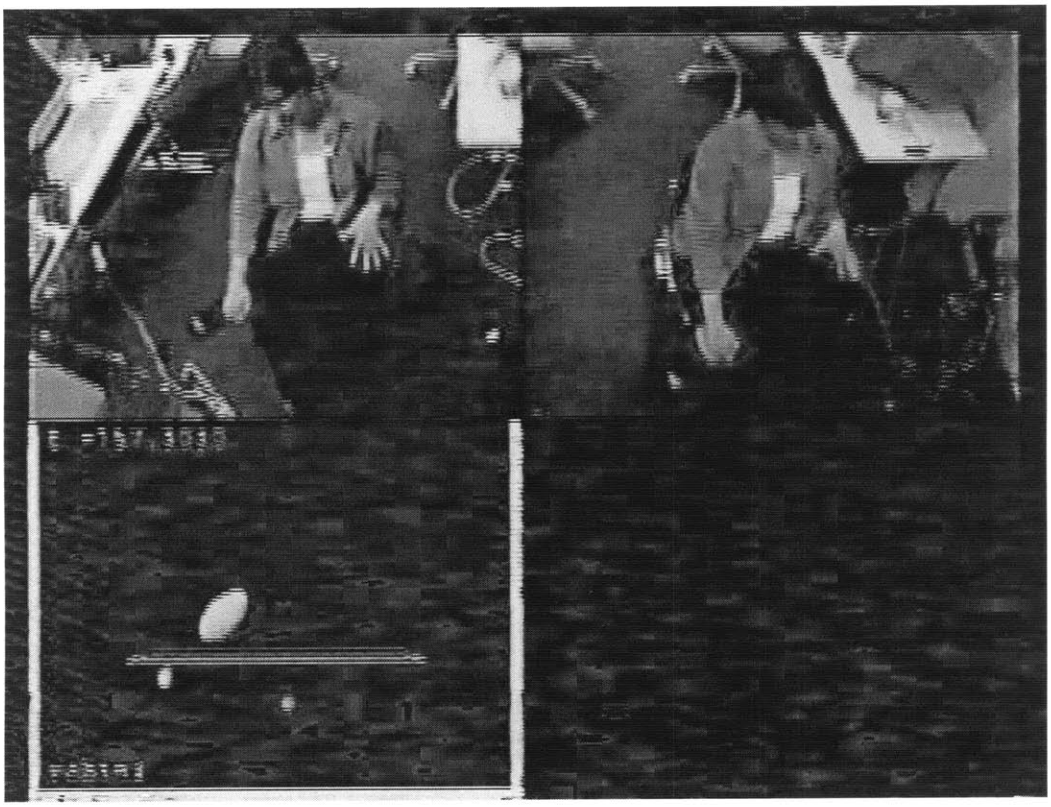

Figure 2-2: Subject RS: "... 70's decor, had like a green shag rug..." Beat shown is on "70's;" left hand flicks partway open. Subject made three subsequent beats on "green shag rug".

Similarly, people have been observed saying "I'll let you know" while making a typing gesture to indicate email; and saying "I was going down the street" while making a steering wheel gesture to indicate driving. In these examples the mode of communication or travel are indicated by iconic gestures. The general problem of understanding iconic gestures is AI complete, but Bolt's system was able exploit demonstratives to understand the iconic gesture accompanying sentences such as "rotate it by this much."

Beat gestures can also carry topic information not present in speech; for example, if the user says "I saw a boy with a dog," a beat may be the only way of determining whether 'boy' or 'dog' is the rheme. Knowing the rheme enables the system to understand what the user considers the new or salient information in the sentence. In figure 2-2, a suject's beats seem to mark the most unpleasant aspects of a house's 1970's decor.

Iconic gestures were interpreted in an extension of Bolt's system developed by Sparrell [50] to understand the gestures accompanying sentences such as "now rotate it by this

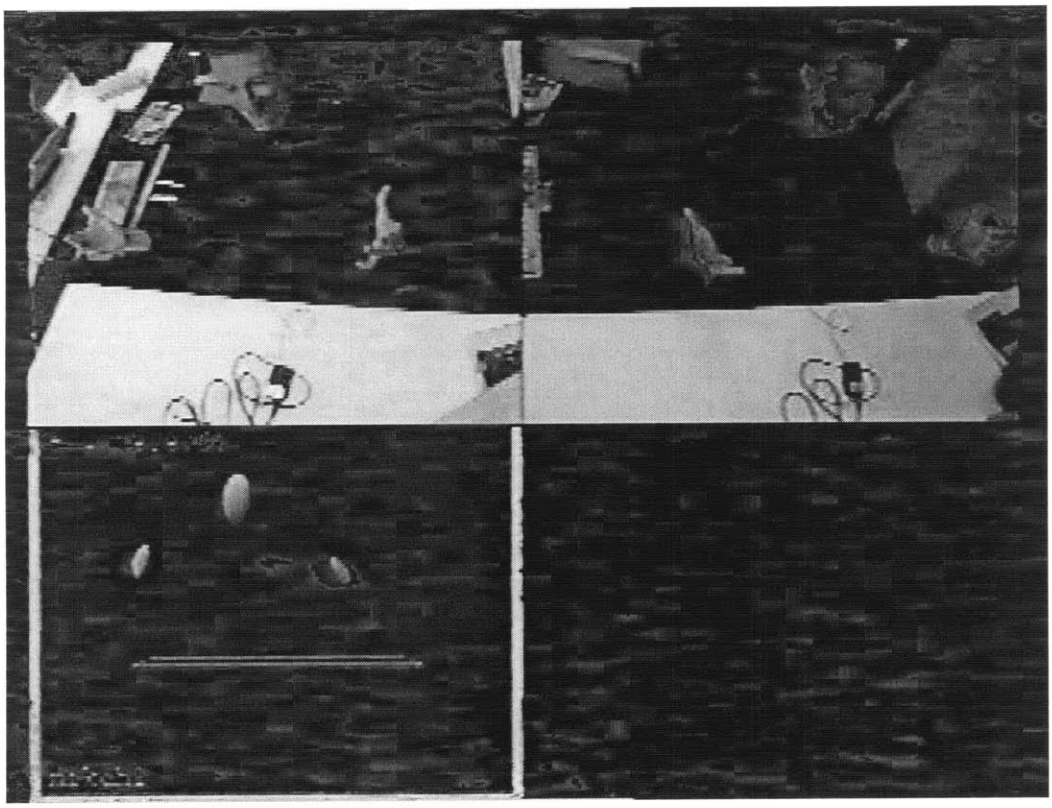

Figure 2-3: Subject MH saying "kitchen was beautiful, it was luminous!" Subject made a two handed metaphoric fountain gesture with his hands palm down sweeping up together, then spreading and drifting down palm up during "luminous".

much," or "extend it this far."[2] Following Sparrell's lead, an automatic system can attempt to interpret iconic gestures when there is a physical preposition (e.g. "put it near that") or an ambiguous physical verb (e.g. "shift it a little"). Iconics and metaphorics may be used as a signal that the topic is continuing (see below, beats and deictics associated with change in topic). In figure 2-3 subject MH makes a metaphoric gesture during an ongoing description of a kitchen.

2.2.2 Interactive gestures

Some gestures do not relate to the topic of conversation, but to the structure of the conversation; these gestures are known variously as regulators, interlocutional, and interactive; and

[2]The system Sparrell and others at the MIT Media Lab Advanced Human Interfaces Group developed was a demonstration disaster recovery system, and it handled commands on placement of firetrucks, digging of trenches, and the like.

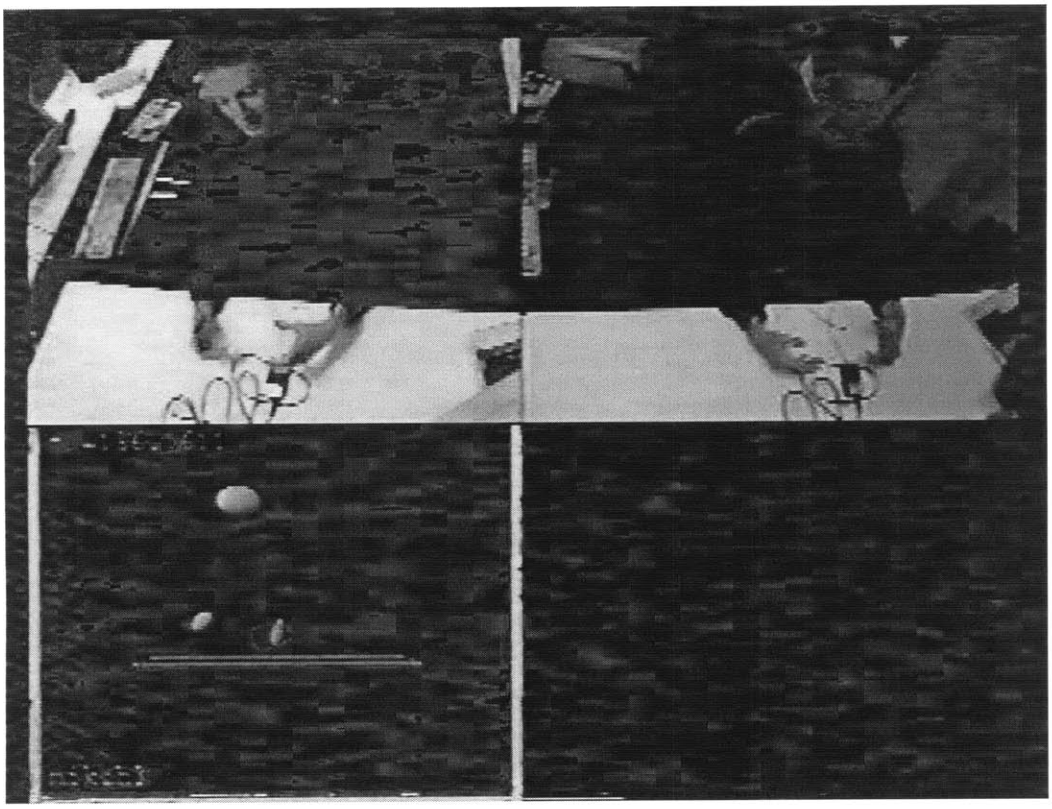

Figure 2-4: Subject MH: "it was nice" followed by preparation gesture. Subject went on to say "some of them were closer together," making a two handed iconic to show relative positions of apartment buildings.

according to Ekman [21], are culturally determined. For example, gesturing towards the other person may indicate that one is done with one's turn; conversely, holding up one's hand may indicate one wants to keep the turn. In figure 2-4, subject MH makes a preparation gesture while thinking about the next phrase he wants to speak.

Bavelas [5] reports four categories of interactive gestures:

- (a) citing another person's contribution;
- (b) requesting agreement, understanding, or help;
- (c) the delivery of new versus shared information;
- and (d) turntaking management.

Bavelas' example of (d) is as a "pushing movement" indicating a speaker's desire to hold the turn or block an interruption, and her examples of (b) are hand flicks towards the listener,

requesting feedback - the gestural equivalent of the verbal tag "know what I mean?" Listeners may also raise one or both hands into gesture space to indicate "I'd like to interrupt." [35].

Beat or Butterworth gestures are often present during repair sequences, which tend to be disfluent, and are thus a useful clue to locating the parsible sections of the kinds of ill-formed statements that commonly occur in real conversation. They also hold the turn in the absence of speech.

Beats and deictics are also associated with the introduction of a new topic by a speaker [34] thus they may serve either content or interactive functions.

2.3 Conversational Characters

Question: even if we are capable of programming a computer to understand these interactive and content gestures, of what use is this feat?

Answer: When the goal is a system to carry on a conversation, the system can gather the extra content information that is present in gestures and not speech, and can react much more naturally by using turntaking and feedback cues and behaviors.

Furthermore, since many people avail themselves of the gestural channel when it is present, a system that is capable of understanding even a subset of gestures will communicate more effectively than one that is unaware of gestures. In cases where the system detects a gesture, but cannot understand it, it is still helpful for the system to reply to the effect that "I could not understand that; please try to explain again."

The REA project [11], mentioned above, is a computer generated character with a graphically generated face and body, that uses various custom and off the shelf software systems to listen to users and observe their gestures, and to speak back to and gesture to the users. REA implements face to face conversation as a computer interface tool. The character REA plays a real estate agent, and operates in a scenario in which she shows the user through virtual houses and condominiums and discusses their features, with the conversational goal of selling the property to the user. Our intention with REA is to exploit as fully as possible the information derived from gesture to make the conversation more pleasant and natural, and to make the user better understood.

2.4 A firmer foundation for gesture classifications

As noted above, it is difficult to discourse research has had a problem regarding the classification of gestures – since classification has always been done by humans while listening to the associated speech, it leaves open the interpretation that for many gestures, speech is primary, and contributes to the understanding of gesture. Krauss et al. [32] addressed the question "Do Conversational Hand Gestures Communicate?" in a series of experiments in which words and their associated gestures were separated, and then subjects were tasked with recovering the association under various conditions. The researchers concluded that gestures were not richly informative. However, their methodology was limited to word level semantics and did not cover other elements of communicative intent such as turntaking, emphasis, and comparison. This thesis will provide a component of the missing proof by inferring communicative intent from the gesture stream, and adding it to the speech stream.

Chapter 3

Context of this thesis

Prior work in the fields of gesture studies, discourse, and machine vision guided the choice of gesture classification scheme, features for recognition, and recognition technique. This examination of prior work puts the thesis project in its historical context and explains the choices made.

3.1 Gesture Classification Schemes

The use of gestures in public speaking and rhetoric has been studied since ancient times – it was discussed by Cicero. However, the spontaneous gestures that people make during conversational speech were not a subject of academic study until David Efron in 1941 [20]. Efron visually classified gestures observed on the streets of New York City using film but not audio recordings.

Efron's goal was to definitively refute Nazi Party claims and theories regarding behavioral inheritance by proving that gestural behaviors are cultural. Along the way he developed three classifications of gestures: (A) Spatio-temporal, i.e. visual or morphologic; (B) Interlocutional, i.e. protocol or interactive; and (C) Linguistic, i.e. content or topic.

Birdwhistell, perhaps the first "kinesicist," in collaboration with linguists and psychiatrists, studied the correlation of gestures with speech beginning in the middle 1950s, with a focus on psychiatry. He associated body motions and styles of speech with categories of

behavior and psychiatric conditions.

Ekman and Friesen [21], in 1969, included gestures in their inventory of nonverbal behaviors. Following Efron, they recognized five categories: beats or batons; ideographic, i.e. tracing out thought patterns; deictic or pointing; physiographic, i.e. representing form or spatial relationships; and symbolic, i.e. the gesture has no morphologic relation to its referent.

Adam Kendon [29, 30], starting in the early 1970s, studied spontaneous gestures in ordinary conversation, and classified them semantically. His early classification schemes were hierarchical, inspired by grammar and linguists, but by 1980 he had developed a much flatter classification similar to McNeill's, which follows.

David McNeill [35], beginning in the late 1970s, also studied spontaneous gestures as a component of language, and made clear that while some gestures have a meaning related to content, others, serve a discourse function such as turn taking.

Table 3.1 summarizes the classification schemes of Efron, McNeill, and Ekman and Friesen. As can be seen in the table, the schemes are quite similar. McNeill's scheme establishes the clearest relationship between the gestures and the discourse functions they serve. Therefore, the functional classification used in this thesis – beats, deictics, iconics and metaphorics – is the scheme developed by McNeill.

3.2 Feasibility of gesture recognition

The research cited above is based on human classification of gestures. This raises the question: to what extent is it feasible for machines to recognize gestures? The following discussion covers work involving automated recognition and / or processing of gestures.

Dick Bolt [8], starting in 1980, and later with the members of MIT Media Lab Advanced Human Interfaces Group [31, 50, 58], developed an automatic system for understanding speech and hand gestures. Speech recognition was based on a commercial product called HARK; gesture recognition they developed themselves using "flock of birds" position sensors and dataglove hand sensors, and a head mounted, corneal reflection eye tracker.

Justine Cassell and her collaborators [12] developed "animated conversation," a system in

which two computer generated characters held a conversation, generating combined speech and gesture. This demonstrated an approach to the problems of generating and semantically representing whole conversations, including the non-verbal components. The two characters avoided the problems of speech recognition by communicating "internally" – textual representations of their speech were passed back and forth between the characters in lieu of a recognition layer.

Kris Thorisson [54] developed Gandalf, a "communicative humanoid" system which answered questions from a human user about planets and the solar system. Whereas animated conversation could use its internal representations to communicate gaze, gesture, and speech between the two characters, Gandalf used Bolt's input system of trackers and sensors.

In a 1992 report on AHIG's work [55] Thorisson et al describe the system as being divided into a map component which displays icons, and an agent component that interacts with a human user. The map displays icons of objects such as trucks, planes, and fire fighting crews. The agent interprets and responds to commands regarding the map such as create, delete, move, name, and request info. If information is missing from a request, e.g. "move *that helicopter* to *there*" the system will attempt to fill in the missing information by considering the point where a pointing finger vector and an eye gaze vector intersect the map. If the agent cannot find a referent from speech, hand, or eye then it will ask the user for clarification.

Bolt and Thorisson use encumbering hand tracking systems that were state of the art at time of use. For improved user comfort and convenience we would like to avoid that encumbrance and use newer visual tracking systems. Cassell's system processed gestures as an essential component part of the discourse, but it avoided the problem of recognition. These next systems address the problem of movement recognition from vision.

Clearly the work cited thus far demonstrates the feasibility of gesture recognition for some classes of gestures. More recent recognition work will be discussed in section 4.1.1.

The next project attempts to understand actions rather than gestures. However, it "understands" them very thoroughly, including cause, effect, and some inferences of intent. It is interesting as a data point indicating just how far a machine can go in understanding some actions given only visual input.

In "Visual Event Classification via Force Dynamics," [49] Siskind presents a system that

converts video input to simple cartoons, and then analyzes the cartoons for the "atomic" relationships of contact, support, and attachment between objects. The program thereby perceives the "atoms" of relationships of contact, support, and attachment from raw input. Next, the program tries to recognize one of 7 simple actions from the sequence of atoms. It can recognize actions like dropping, picking up, putting down, carrying, and stacking. Thus it's notion of the meaning of a word like "stacking" is as a sequence of changes in atomic relationships between objects.

The system looks for lines in a scene, and groups those lines into objects by inferring which lines are rigidly attached to each other. It makes guesses about rotating and sliding joints between objects, and which objects might be "grounded."

The system has a 2 dimensional layered view of the world, and makes inferences about whether objects are on the same layer by whether they appear to contact or "pass through" each other.

It then enumerates all stable interpretations of the scene based on the minimal sets of grounded, jointed, and rigid relations - these are minimal models of the scene, and prefers the model with the fewest assumed relations.

Finally, it examines frames and models over time and prefers the sequence of models that entails the fewest changes in assertions.

The system has a relatively crude vision system, and requires a uniform background and objects with simple contrasting colors. Given those limitations, the system is very effective at classifying motions involving a hand playing with blocks. The most interesting thing about the system is its construction of meaning from perception through "atoms" into motion verbs. I am aware of no other system with as complete an understanding of motion verbs.

Unfortunately no one has yet devised a way to reduce co-verbal gestures to a small set of perceivable atoms. The meanings of co-verbal gestures are dependent on the co-occurring speech, and speech has resisted many attempts at reduction.

3.3 Features for recognizing human motion

The perceptual experiments of Johansson in 1973 [27] opened up the field variously known as "Biological Motion" or "Moving Light Displays" or . Johansson attached reflective patches to ankle, knee, hip, shoulder, elbow, and wrist joints, and to the head, and took videotapes adjusted such that only white spots on a dark background were visible. In some experiments, subjects viewing the videos were able to tell whether or not the motion was biological in times as short as 100 or 200msec [28]. In other experiments, common motions were added to or subtracted from all the points, and subjects still identified the motion. In most of the experiments 100% of the subjects responded correctly, even though they had not been prepared or trained in any way. Another experimenter [19] found that other activities were recognized such as hammering, lifting a box, bouncing a ball, and stirring; and two-person activities such as dancing, greeting and boxing. Yet another series of experiments [4] showed that people could determine the gender of a walker and even identify a friend based on 'Moving Light representations of their walk.

Johansson proposed a rough model which he called "visual vector analysis." The model says that when a velocity vector can be abstracted from a group of objects, the objects will be perceived as being in unitary motion. However a comprehensive algorithm is not given.

It is surprising that people can make such quick identifications from such a paucity of data, and that the effect is so robust. For purposes of this thesis, the implication of Johansson's work is that a small number of point motions (such as that produced by a hand tracker) may be sufficient model for many motion recognition tasks.

Johansson's inspired a number of early vision researchers, who began by attemting to recognize mechanical, as opposed to biological motion.

Rashid [44] made an early attempt to link points into a structure based on clustering of 2D positions and velocities. His system tracked points through frames, grouping and linking them into objects according to their relative velocities. The system was tested on synthetic 2-D perspective projection MLD's of several objects such as a man walking a dog. The points that were rigidly linked in space as well as independently moving sub-parts could frequently be found by the program even though their projections were not rigidly linked. Although

velocity clustering may provide a good first estimate of how points are linked into objects, it is too weak a model of objects to be relied on.

Webb and Aggarwal [57] assume the links between points are known, and that points rotate around fixed axes. They use orthographic projection and show that points will sweep out ellipses in the projection, and that the eccentricity and orientation of the ellipses determine the 3-D axis of rotation to within a reflection. Their work was later extended by Asada [1] (with an assumption of constant angular velocity) to a case where, for example, points rotate about the shoulder which in turn rotates about some other axis.

Hoffman and Flinchbaugh [25] present a method which uses the fixed axis assumption of Webb, but recovers both linkages and axes. Thus they recover 3-D structure in the case where the axis directions are fixed. This is a good approximation for the cases of walking and running, but not dancing.

None of these three recognition systems started with images – they all began with sequences of moving points. They all make similar assumptions about the input: that points or features corresponding to limbs can be found and tracked through successive frames to provide an MLD input. Webb assumes starting identification (i.e. linkage) is given; Webb and Hoffman assume fixed axes. This body of work, taken as a whole, suggest tracked points are a reasonable set of inputs for a recognition system. The next two systems use visual input and view based motion features.

Quek and McNeill [40] use view based positional features, essentially pixel addresses, as features for hand tracking. These features are scrutinized by human analysts as part of a by-hand video analysis process. They report no automated recognition of gestures. In a newer work, Quek and McNeill [41] use a 3D hand tracking system similar to STIVE except that it does not do online tracking. Cartesian positional features are derived and scrutinized by human analysts as part of a by-hand video analysis process. Attention is drawn to regions of steep slope in the cartesian plots; essentially these are regions of higher velocity where velocity is estimated by viewing the slope. Thus it can be said that Quek's human analysts are using velocity features. No automated recognition of gestures or catchements is reported.

Davis and Bobick [7] developed a realtime view based approach to human motion recognition that makes use of two images derived from a sequence of video. The two images are

the motion energy image, a binary image, and the motion history image, a grey level image. Scale and translation invariant Hu moments are calculated from the MEI and MHI, and the moment coefficients are compared against those of exemplars to recognize a movement. The technology was demonstrated in a system called PAT, the Personal Aerobics Trainer [17], and in a children's interactive playspace called the KidsRoom [6]. An advantage of this approach is that only one instance of a movement is needed as an exemplar.

A related approach to realtime view based human motion recognition was developed by Cutler and Turk [47], who derive optical flow from a sequence of images, cluster regions of similar flow, and fit ellipsoids, referred to as "blobs," to the clusters. Sizes, velocities, aspect ratios, and qualitative motion parameters are then computed for the two largest blobs, such as up, down, left, right, and rotating. Subsets of these parameters are then used to directly classify movements, for example clapping is recognized as two blobs in horizontal motion with opposing directions. The system recognizes six movements, and once a movement is recognized it estimates frequencies of oscillating parameters. The approach is incorporated into an interactive environment for children. There is an opening screen which show the various movements and what they control – for example there is a conducting movement in which the speed of the movement controls the tempo of a song.

Both Cutler and Turk's system and PAT have the advantage very simple training procedures. They are also both able to take advantage of their application domains by controlling the location and orientation of the user, and by choosing a set of well defined movements easily differentiated by their methodologies. However, these two systems necessitate view based features. As will be discussed, tracking systems are more robust at ignoring of background motion, and body centered features have advantages over view based features.

3.4 Difficulties of co-verbal gestures

Although there has been much work done on automated gesture processing (for reviews see [13, 26, 38, 23]), almost none of it addresses the problem of recognizing co-verbal gestures. The two main thrusts of existing research are (1) measuring gestures with respect to a model in order to manipulate a virtual object, where the identity of the gesture is implied by

the grammar of the manipulation system; and (2) recognizing gestures from a well defined set with clear notions of proper formation.

Co-Verbal gestures are more problematic. Approach (2) may not be straightforwardly applicable because there are not right and wrong ways to make co-verbal gestures; whatever it is that makes a beat distinct, it's more abstract than the kinds of features common in the literature. Meanwhile, approach (1) presupposes some additional source of information to supplement the gesture measurement – but we wish to do the reverse and use gestural information as an independent source to supplement speech. Some of the ways in which co-verbal gesture classification can supplement speech are enumerated in Section 6.10

Another difficult issue is bridging the divide between functional and morphologic categories. The morphologic categories are what are more easily seen by a computer with a camera; the functional categories are of more use in understanding the communicative intent of a gesture.

3.5 Functional classifications

Table 3.1 is a comparison of the three major functional taxonomies of co-verbal gestures. These investigators do not disagree on the broad divisions; only on the subdivisions within iconics and metaphorics. These subdivisions are semantic, rather than visual or functional, and thus cannot be made without understanding the associated speech.

3.6 Morphologic classifications

Efron's spatio-temporal aspect of gestures was the first attempt at making morphologic (shape and pose) distinctions. Projecting 16mm films on graph paper, he recorded the radius, axis (wrist, elbow, etc.), form (sinuous, elliptical, angular, or straight), plane, body parts involved, and tempo of gestures. However, he did not develop morphologic categories but rather used his feature set to demonstrate cultural differences between groups of gesturers. Koons [31], Sparrell [50] and Wexelblat [58] used morphologic features such as hand poses and path angles to identify deictics and a subset of iconics. Cassell [12] identified the morphologic clas-

McNeill Categories	Efron Categories	Ekman and Friesen Categories
Iconics	Physiographics, Kinetographics	Kinetographs, Pictographs
Metaphorics	Ideographics	Ideographs, Underliners, Spatials
Deictics	Deictics	Deictics
Beats	Batons	Batons, Rhythmics
Butterworths		

Table 3.1: Comparison of gesture taxonomies, from McNeill.

sifications of "biphasic" and "triphasic," where biphasic contains two motion segments, and includes beats, deictics, and butterworths, and triphasics have three motion segments and include iconics and metaphorics.

3.7 Lack of computational gesture recognition models

To date, there has been almost no work on computational models of co-verbal gesture recognition. Wexelblat [58] worked towards that goal, converting tracking information and dataglove data into a representation designed for recognition, but it was only applied to "put that there" and "move it like this" tasks. The lone computational model of co-verbal gestures [61], uses principle components analysis to derive view based features and classify gestures into the categories of biphasic, triphasic and resting [10]. They used a Markovian state machine with explicit duration modeling for the biphasic and triphasic gestures, and they used a correlation distance metric to identify the most commonly repeated subsequence as a model for rests. The results were presented graphically: not per gesture, but per second of elapsed time, thus they are difficult to compare to other recognition systems.

3.8 Gesture occurrence frequencies

Based on data in [35] Table 3.7, 45% of gestures are beats; 45% iconics; and 5% each are deictics and metaphorics. This data comes from an experimental task in which subjects describe the action in a cartoon. It is further subdivided into gestures during narrative clauses, i.e. describing events in the plot; and extranarrative, e.g. describing the setting, introducing the characters, etc. In extranarrative speech, beats make up 66% of the gestures, iconics 17%, metaphorics 15%, and deictics 2%.

For this thesis I gathered gesture data by conducting interviews about real estate, in which subjects describe where they live and work. This data shows gesture frequencies of about 24% beats, 38% iconics, 6.2% deictics, 12% metaphorics, 9.3% preparations, 8.9% retractions, and 1.5% butterworths. My methodology differs from McNeill's in that every gesture begins and ends with a period of zero hand velocity. Thus if the preparation or retraction phases are continuous with the gesture, then they are not segmented as separate gestures. By ignoring for a moment the preparation, retraction, and butterworth categories and renormalizing, my proportions come closer to McNeill's except that I observe more deictics and metaphorics. The difference in domains may explain this discrepancy in gestures frequencies.

The majority of real estate related gestures were beats, iconics, and deictics; therefore, in order for REA to understand discourse in the "real estate" domain it would be useful to be able to recognize those three categories of gestures. It will also be useful to be able to reject preparations and retractions so as not to confuse them with other gestures. Since iconics and metaphorics only differ semantically, and are indistinguishable morphologically, they will be merged into a single category for recognition purposes.

3.9 Approaches to Visual Gesture Recognition

A number of approaches to gesture recognition have been developed over the years by the machine vision community. However, it is important to note that the gestures being recognized are usually a set of 20 or fewer gestures and static hand poses defined for a particular task, or else a subset of a sign language. Both sign language and the task specific gestures

tend to be well-formed, and thus recognition of these gestures is more tractable than recognition of co-verbal gestures. With this in mind, here is a sampling of other approaches to visual gesture recognition. The systems by Davis and Bobick [7, 17] and Cutler and Turk [47] have already been discussed in Section 3.3.

Quek and Zhao [42] developed a system to recognize a set of 15 hand poses and gestures "designed for the description of space and the specification of spatial quantities." In their system, the hand is placed in front of the camera such that it nearly fills the camera view, and the background is black. From each frame, a set of 28 pixel-based features is computed, such as area of bounding box, principle axis, and normalized moments. The system trains on example images of poses. For each pose, the system automatically develops a set of rules involving the 28 features that discriminates the pose from all other poses. Although the automatic rule learning is impressive, Quek and Zhao's approach is not amenable to co-verbal gesture recognition for three reasons: (1) it requires images filled by the hands, i.e. a full body shot would not work; (2) it expects a particular view of the hand pose, so it would have trouble if the user rotated; and (3) it expects a set of well-defined gestures.

Cui and Weng [15, 16] recognize a set of 28 American Sign Language signs using maximally discriminating (MDF) image weights as features, and a vector quantization of a low dimensional MDF space as a distance metric. Their system is more suited to recognize hand poses than hand movements, and so it is instructive to compare it with the HMM based approach of Starner and Pentland [52], who also recognize ASL signs. Cui and Weng report a 93% recognition rate on 28 signs using an offline system; Starner and Pentland report 99.2% recognition on 40 signs in realtime. Part of the reason for Starner's very high recognition rate is the use of grammatical constraints in his ASL sentences; even without grammar he reports 97% correct recognition. Since HMMs are both much faster and more accurate than Cui and Weng's system on ASL, it is reasonable to conclude that HMMs will perform better on co-verbal gestures.

Cohen et al. [14] developed a gesture recognizer for a set of 24 "oscillating motion" gestures such as those used by construction workers to pass signals to crane operators. The gestures include fast and slow horizontal, vertical, and diagonal movements, and large and small circular movements. Recognizers are hand-built dynamic models or differential equa-

tions which represent different trajectories. The models effectively capture independent features of the motion, such as sinusoidal vertical, horizontal, in-phase, and out-of-phase. The set of oscillating motion gestures is well matched to the features, but would not work well with the non- sinusoidal motions encountered in co-verbal gestures.

Gutta et al, [24] recognize a set of 25 hand poses using a recognition system that involves neural nets, radial basis functions, and decision trees. The system trains and tests on normalized images of hands that fill the viewframe. The authors note that their system "assumes that hand gestures have already been located and normalized." Thus it requires some other input to tell whether or not a hand gesture has been performed before it can do its classification. There are two reasons why this method won't work well for co-verbal gestures: (1) like Quek's system it requires images filled by the hands, and (2) it doesn't know when or if a gesture is happening.

Triesch and Marlsburg [56] recognize a set of ten hand poses by representing each hand pose as a bunch-graph of 35 nodes and 70 edges, and elastic matching of graphs. The nodes of the graph are values from human-visual-system-inspired Gabor filters. The system is quite robust to changes of background (as long as the background isn't another hand!) but it expects a set of well-defined poses, and thus is not suitable for co-verbal gesture recognition.

Pavlovic [37] compare HMMs with Switching Linear Dynamic Systems to recognize fronto-parallel walking vs jogging. Input is joint angles, fit by Rehg's [45] tracker. They found SLDS had better segmentation accuracy than HMMs on their task. However, the published graphs seem to indicate SLDS's produced more glitches than HMMs, so the comparison is inconclusive. Four state model HMMs and SLDSs did the best and produced very similar results.

Chapter 4

Recognition / classification methodology

4.1 Why use Hidden Markov Models for gesture recognition?

An ideal gesture recognizer would be AI complete (i.e. as smart as humans, combining speech and gesture at multiple levels of understanding). Since this is impossible, this thesis will classify gestures based on morphologic characteristics. A good morphologic classifier would have these desirable properties:

- Invariant to changes in viewpoint of camera
- Invariant to translation and rotation of user
- Able to model the "typical variation" in gesture performance
- Robust to "small variations" of gesture speed and shape

HMMs were chosen because they provide the last two properties, as will be seen below. In addition, HMMs are able to segment gestures out of a stream of data – a significant advantage because otherwise a gesture recognition system needs a separate segmenter. If the system must run in realtime, this puts even more demands on the segmenter.

Invariance to camera viewpoint is provided by STIVE [2], a hand and face tracking system used in this thesis and described more fully in Section 5.3. Since STIVE measures head and hand positions in three dimensions, measurements can be translated into a user centered coordinate system independent of camera location. In theory, as long as the cameras can continue to see the user's head and hands, camera location is irrelevant. In practice, there is measurement error (noise) in STIVE measurements, which is nonlinearly dependent on camera position, and a wide baseline helps reduce measurement noise.

Invariance to translation and rotation of user is provided by choice of feature [9]. For example, using velocity features yields translation invariance; if the velocities are in polar coordinates (with the head as the pole) there will also be rotational invariance about the users head. Correlation features are also invariant to rotation and translation.

Hidden Markov Model recognizers excel at modeling typical variation, and thus are robust to small variations. They also excel at finding the nearest match when data is missing or extraneous data is inserted. They were first applied to speech recognition with great success in the 1980s [43]. HMMs are used to recognize phonemes, words, and even sentences. They are now being used in genomic research to search for similar genetic sequences because of their robustness to insertions and deletions (continuous HMMs are used for speech recognition, while discrete HMMS are used for genomics).

The HMMs used in this thesis are trained on examples of naturally performed gestures. Presumably the training set captures a range natural variation both within speaker and between speakers.

Forward chaining continuous HMMs are appropriate to recognition of well-formed gestures because gesture performance and measurement is a stochastic process: there is an underlying deterministic intent to make the gesture in a particular form; added to that is actuator noise (muscles are imprecise) and sensor noise (the stereo vision system has estimation error).

Co-Verbal gestures do not have standards of form (e.g. compared to sign language), but perhaps the gesture categories do have some regularities of timing and form which can be exploited by HMMs. Therefore, this thesis will consider the claim that, like the more linguistic gestures, there is some underlying deterministic intent in a co-verbal gesture, and additional variation can be modeled as noise. Thus, some function $\mathbf{f}(x, y, z)$ of the sequence

of coordinates of the hands, as sampled over time and measured by the vision system can be considered to be the observable symbols of a Hidden Markov process.

4.1.1 Features for well defined gestures

Past uses of HMMs in machine vision used a variety of features. Yamato et al. [63], probably the first to use HMMs in machine vision, approached the problem of recognizing tennis swings, using as features a 25×25 pixel binarized camera image. This approach to recognition depends on an unchanging background and the same camera viewpoint. Schlenzig et al. [48] used HMMs to recognize gestures in sequences, using a rotation invariant representation of a binary image, processed through a neural net. This approach also requires a constant background. Wilson and Bobick [60] incorporated multiple representations in an HMM framework, using eigen image weights as features. This approach is viewpoint dependent.

In more recent work, Wilson and Bobick [59] developed Parameterized HMMs (PHMMs), and used STIVE (see Section 5.3) as input, overcoming viewpoint issues. PHMMs allow a free vector parameter, $vec\theta$, to be present in both the training and test data. The free parameter is estimated as part of the HMM recognition process. The linearity constraint can be relaxed by choosing $vec\theta$ to be a nonlinear function of features. This is very promising work, however it has not yet been demonstrated in a realtime system.

Starner and Pentland [52, 51] used HMMs to recognize 40 American Sign Language gestures, grouped into 5 word sentences, in real-time 2D video imagery, tracking the hands via skin color, using the 2D image coordinates and orientation of the hands as feature vectors. The first to use HMMs to recognize ASL, they achieved recognition rates as high as 98% when using grammatical constraints, and 92% with no grammar.

The T'ai Chi project in which I was involved [9] was based on Starner and Pentland, but with a 3D tracking system that would allow a more robust approach to the problem of user translation and rotation. We selected a set of 18 T'ai Chi movements, and created random groups of six movements analogous to Starner's sentences. We used STIVE to track the hands in 3D, and examined the recognition rates of different features derived from body centered coordinates extracted from the tracking data. We achieved recognition rates as high as 95% for the best feature set.

Pavlovic [39] developed a hybrid system that essentially placed Kalman filter between the input data and the HMM. The Kalman filter can be used to regularize the input features or estimate unobservable features. They demonstrate a 95% recognition rate on a set of 4 mouse gestures. A special case of KF derived features is work by Wren [62] which trained HMMs on Kalman Filter innovations to recognize a set of 3 gestures. Innovations capture the parts of the movement that the KF is unable to predict. Although this idea is quite interesting, it was only pursued to preliminary and unpromising results.

Hidden Markov models achieved the greatest increase in recognition rates over random chance compared to other approaches to machine vision recognition of human movements. The ASL project achieved 98% where random chance would have been 2.5%; the T'ai Chi project achieved 95% where chance would have been 5.6%. Since HMMs demonstrated the best ability to recognize, they were selected as the recognition method for this thesis.

4.2 Choice of HMM features

Having selected HMMs as the recognition method, we now turn to the question of a feature set. To motivate the choice of feature vectors, consider a person sitting at a fixed location and orientation with respect to the vision system, and performing a set of gestures, and assume the HMM trains on and recognizes (x, y, z) coordinates of hands. Then the states of the continuous HMM will correspond to density functions in space that the hands pass through. If the hands move faster or more slowly during a test gesture, the Viterbi algorithm will compensate for this; effectively performing dynamic time warping. However, if the gesture is made "smaller," it will only pass through the fringes of the density functions and will receive a lower log-probability score. Also, this feature is not invariant to translation or rotation.

Now consider the velocity of the hands (dx, dy, dz) as a training and testing feature: states of the HMM will correspond to velocity vectors at different times during the gesture. This feature is translation invariant but not rotation invariant. It also gives up some of the dynamic time warping ability of the Viterbi algorithm: if, during a test gesture, the hands move at lower or higher speed, the symbols will fall more to the fringes of a state. However, if the hand moves at the same speed for a longer or shorter period (i.e. makes a bigger or smaller

version of the gesture), time warping can occur.

Ideally, a set of features should be shift and rotation invariant; isotropically noisy in the measurement space; able to make use of Viterbi time warping; and contain as much "context" as possible (i.e. values unique to each stage of a gesture). Unfortunately, these ideals cannot all be fulfilled at once; they trade off against each other. For example: imagine moving your hand in a perfect circle. Curvature and speed (ρ, ds) will be completely invariant to rotations and translations but will be constant and thus exhibit no "context." At the other extreme, (x, y, z) coordinates can easily distinguish the top, bottom, and sides of the circle, but they require it to be performed each time at the same location. In between these extremes, (dx, dy, dz) and $(dr, d\theta, dz)$ trade off some invariance for some context. An additional problem with derivatives is that they amplify high frequency noise.

This problem is present in the curvature features because curvature is a function of the second derivative and thus inherently more noisy than velocity features. A few tests using acceleration as a feature showed that second derivative noise was a hindrance to recognition. Another problem with curvature is it does not fall in a Gaussian distribution which the HMM expects; particularly when the hand is almost stationary it tends to generate very high curvatures and very low velocities. Taking $\log(\rho)$ or $\log(\rho ds)$ shapes the distribution to be more Gaussian and improves HMM recognition rates.

Including dz as a feature means that at best we can achieve invariance to rotations in the horizontal plane, but not to out of plane rotations. However, people tend to orient themselves with respect to gravity, so out of plane rotations will be unusual.

The result of the T'ai Chi study was that, as one would expect, features designed to be invariant to shift and/or rotation perform better in the presence of shifted and rotated input. Cartesian velocity features perform better in the presence of translational shifts; polar velocity features perform better in the presence of rotation. Also, it became clear that choosing the right set of features can be crucial to the performance. There is a design choice that goes into every implementation of gesture recognition systems and since this choice can greatly impact performance, it is critical that one understand the effect of feature choice on performance. Finally, given a finite training set, too many features will produce an HMM with too many parameters, resulting in overtraining and degraded performance on real world data. Thus

it is critical both to be able to detect the point of overtraining, and then keep the feature set small enough to prevent it. This will provide a constraint on the number of features; then the features themselves must be selected carefully.

The gesture classification method used in this thesis is Hidden Markov Models, with a feature vector ideas derived from the features tested in a T'ai Chi movement recognition project [9], and training data extracted from hand-annotated data currently being collected and annotated.

4.3 Feature Vector considerations

In the T'ai Chi study, feature vectors were kept at 3 elements per hand because a fixed length feature vector simplifies direct comparison of feature vectors by maintaining the same number of HMM parameters across all tests; thus all models will be equally over- or under-fitted in terms of parameter count.

4.3.1 Vector Length

A reason to keep feature vectors short is to prevent over training, because when using full covariance matrices, the number of HMM parameters can grow large rapidly. A quick calculation will illustrate: in the T'ai Chi study there were three features per hand, two hands, and five HMM states per model (gesture category). Thus for each state there are six parameters to represent the mean, and 21 more to represent the covariance, yielding a total of 135 parameters per model. If we add one more feature per hand, there will be 44 parameters per state, and 220 per model. The more parameters in a model, the more training data is required to estimate the parameters. In the above example, a single Gaussian per HMM state was assumed – never a mixture of Gaussians. A mixture of Gaussians would allow the system to better model non- Gaussian distributions and to allow multiple paths through a model - effectively automatically learning multiple forms of a gesture. But if a mixture of m Gaussians were to be used, it would increase the number of parameters by a factor of m.

Although there is no simple way to compute how many training examples are needed to accurately estimate N HMM parameters, we can detect the onset of over-training by com-

paring the disparity of results of models with different numbers of HMM parameters trained on separate data sets by this rule: as long as an increase in the number of HMM parameters per model produces higher classification accuracy when training and testing on separate data sets, then there is adequate training data, and overtraining is not yet a problem. Using this test, the feature vector can be extended to be as long as the training data permits.

4.3.2 Vector Contents

Since training data is difficult to obtain – finding the start and end times of each gesture is a slow and painful by-hand process involving multiple slow motion viewings of each gesture – additional features must be added parsimoniously; thus the natural question is "which features?" Among features tried in the T'ai Chi study, velocity features produced the highest accuracy of classification; they are good candidates. Velocity features also make intuitive sense because they are shift invariant and robust to small rotations. Also in the T'ai Chi study we found that features representing vertical components of the motion yield better classification than other components; thus extra Y features are logical additions – for example Y acceleration and velocity, or Y velocity and position (adjusted for head height).

It may also be useful to have some features representing how "repetitive" a motion is, in order to better distinguish butterworths from beats, and possibly to detect either / or comparisons. I developed an "autocorrelation" features in which the most recent e.g. 300msec window of the gesture is compared with a series of earlier windows shifting backwards in time for about a second. One would expect the present window to be most similar to windows nearby in time that strongly overlap it; so that a plot of correlation vs time shift should decay with increasing time shifts. However, if the motion is repetitive over the timescale of the window shifts, an additional peak should occur in the plot. The height, and perhaps timeshift, of the second peak should make a feature representing repetition.

Another approach is to treat the two hands entirely independently. In the T'ai Chi study, feature vectors contained six elements; three from each hand. Gestures were subdivided into "left," "right," and "both;" thus there were three kinds of beat, three kinds of illustrative, etc. If the hands were treated independently, using features that don't discriminate between the hands, there would be no need for this subdivision into three, so more training data would be

available to each model; and simultaneously the length of the feature vector would be cut in half. There are costs associated with such a simplification: any corroborating evidence contributed by the other hand would be ignored; e.g. if most iconics were two-handed, this cue would not be used. Also, features would be restricted to ones that are independent of hands; e.g. if the left hand tended to move slightly leftward on a downstroke, and the right hand rightward, this cue would be lost when the two models merged. However, the shortening of the feature vector might reduce parameter count enough to allow a mixture of Gaussians; thus compensating somewhat for the loss of distinguishing cues in the merged models.

Still another approach is to try to use different feature sets for different models. This brings up the problem of how to compare results from models trained on different feature sets. For example, if feature set A says the gesture was most likely a beat, while feature set B indicates deictic, we need a way to arbitrate between the two choices – a straightforward comparison of log probabilities will not work. All the features could be combined into a longer feature vector, however the longer vector would demand more training data.

4.4 Training Data

Unrehearsed, spontaneous gesture data was gathered from "naive" subjects as they spoke and gestured to an interviewer over a video conference system in which the subject can see a life-sized image of the interviewer. This data was used offline to select features and train the HMMs for real-time classification. Figures 2.2.1, 2-2, 2-3, 2-4, 4-1, and 4-2 show video frame grabs of the subjects in the data collection system. The two views in the upper quadrants are from the two cameras of the tracking system. The three blobs in the bottom left quadrant show the output of the tracking system. The bottom right quadrant is blank.

Subjects were pre-selected for gestural expression (although we have interviewed some subjects who kept their hands flat on the table at all times, we got no data from them that might contribute to the classification system). Subjects were told that they are taking part in a discourse research project, and that they will be interviewed via a teleconferencing system about real-estate related topics. Further details are left vague until the end of the experiment, when the subject was debriefed and the gesture tracking purpose explained.

Figure 4-1: Subject DC saying "... with a team of like real architects they got together ..." with a two handed iconic on "got", hands placed together showing the architects together.

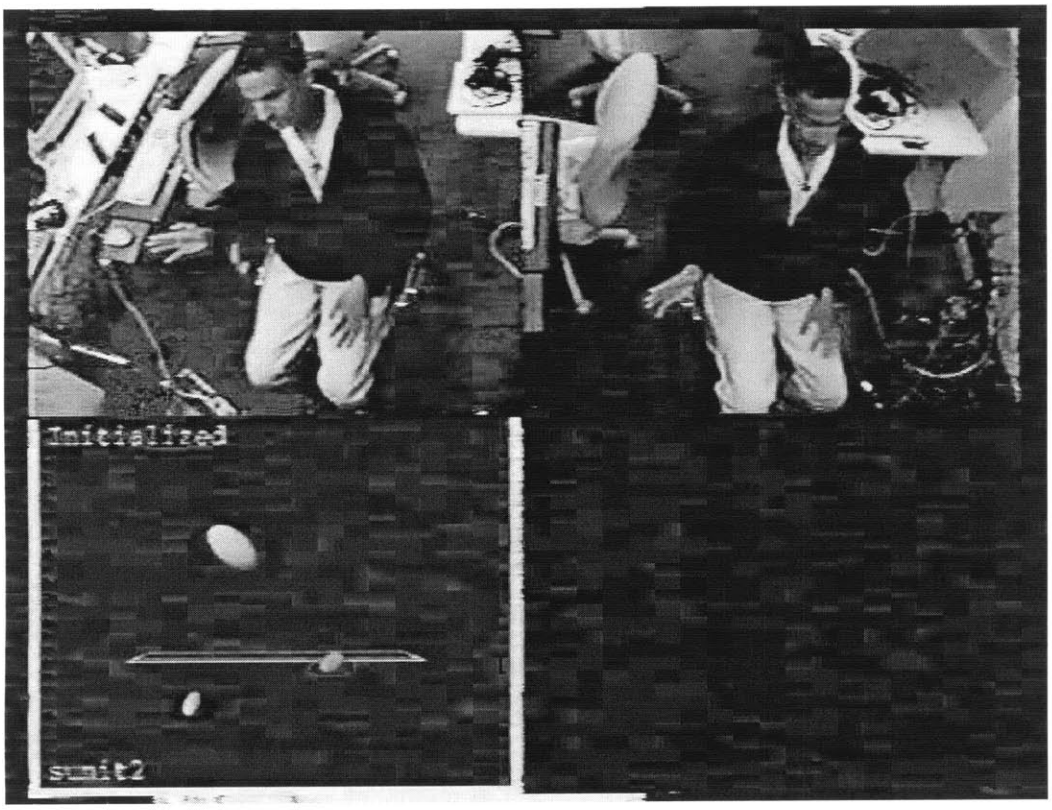

Figure 4-2: Subject SB saying "... the livingroom was on the right..." while flashing his right hand out to the right during "right".

Subjects were interviewed for 10 to 20 minutes on subjects relating to houses, apartments, workplaces, etc. Usually the first three to five minutes contain stiffer or more inhibited gestures. Although I attempted to annotate everything, hand clasping and adjustors caused much of these first few minutes to be discarded. When subjects appeared sufficiently warmed up we asked them to describe the best place they have lived; the worst place they have lived; where they grew up, and similar questions. As they talked, we asked them to describe the layout in more detail. It also was helpful for the interviewer to gesture; it seemed to give "permission" to the subject to do likewise.

The video conference display where the subject sees the interviewer is part of a STIVE system, and as the subject speaks her hand and head data are tracked and recorded. In addition, the two STIVE camera views and a computer graphics view, generated from STIVE data, showing STIVE blobs and timestamps, were all fed through a video quadsplitter and recorded on videotape along with the audio track of the interview.

For annotation, these tapes were played, and every hand movement is labeled, along with its start and end times as defined by a period of zero velocity. This annotated movement data was used to train and test HMMs to develop feature vectors, and ultimately to train the HMMs that are used in REA's classification system.

4.5 Number of Classification Categories

For output gesture categories, we will base our choices on the system of McNeill [35], who finds five categories: iconics, metaphorics, deictics, butterworths, and beats. However, we will accept from the start that iconics and metaphorics are not visually distinguishable, so the two will be merged into one output category called "illustrative".

However, we may need additional internal categories: for example, the system needs to know when the user is not gesturing. A straightforward way to determine this is to have a model for resting. There are many other movements the user can exhibit such as scratching (adjustor) or saluting (emblem). Some of these come up so rarely that it is impractical to create a new model for each one.

Instead a practical way to deal with them is to have one more model for all of the "unclas-

sified" movements. To see how this might work, think of the distributions associated with each HMM state for each model. They are Gaussian blobs or mixtures of Gaussians, all with some covariance. The "unclassified" model will be the same, and if the training data tend to give its states a larger covariance than the other models, then it will effectively determine a set thresholds around the other models' states. Thus it will capture outlying data points using the standard mechanisms of HMMs, without any artificially chosen thresholds.

Additional internal categories may arise through the mechanism mentioned earlier, of subdividing each category into "left," "right," and "both hands" models. These subdivisions would not need to propagate to the final output categories because the output categories represent discourse functions, which are independent of the handedness of the gesturing.

Finally, we may notice subdivisions within the categories and choose to manually create subdivisions to improve accuracy. For example, we may notice distinct horizontal and vertical beats, and create two models trained on their different characteristics.

However, the most likely case is that although there exist subdivisions within some of the gesture categories, we fail to notice them and explicitly create separate models for them. Even in this case, the machinery of HMMs can compensate and learn multiple models via skip transitions. The existence of skip transitions in the HMM topology creates multiple paths through the HMM, and it is possible for two different paths to train on the two different subdivisions within a category. Also, if the data permits using mixtures of Gaussians in each state, it is possible for the different Gaussians to cluster around the different subdivisions.

Chapter 5

Implementation

The gesture classification system described in this thesis acquires input from the STIVE system and generates output for the REA system. The REA system is a collaborative project under continuous development by the Gesture and Narrative Language group. In this section I provide an overview of the REA system, and details of the implementation of the gesture recognition system and the data collection and training system used to develop the recognition models. The description will begin with REA, and then follow the flow of data in the recognition system, starting with training the HMMs and ending with REA's gesture recognition system.

5.1 Description of REA

The REA project is a testbed for experiments in multi modal, mixed initiative conversational interaction between human and computer. We plan to explicitly test theories about the conversation among humans, and between humans and computers. The domain of conversation is real estate – REA shows the user around a virtual house. REA is displayed as a synthetic animated female real estate agent.

REA's hardware currently consists of a 60 inch (diagonal measurement) high resolution rear projection video display, a head-mounted noise cancelling microphone, two color video cameras aimed at the sensing area in front of the display, two SGI O2 workstations to process

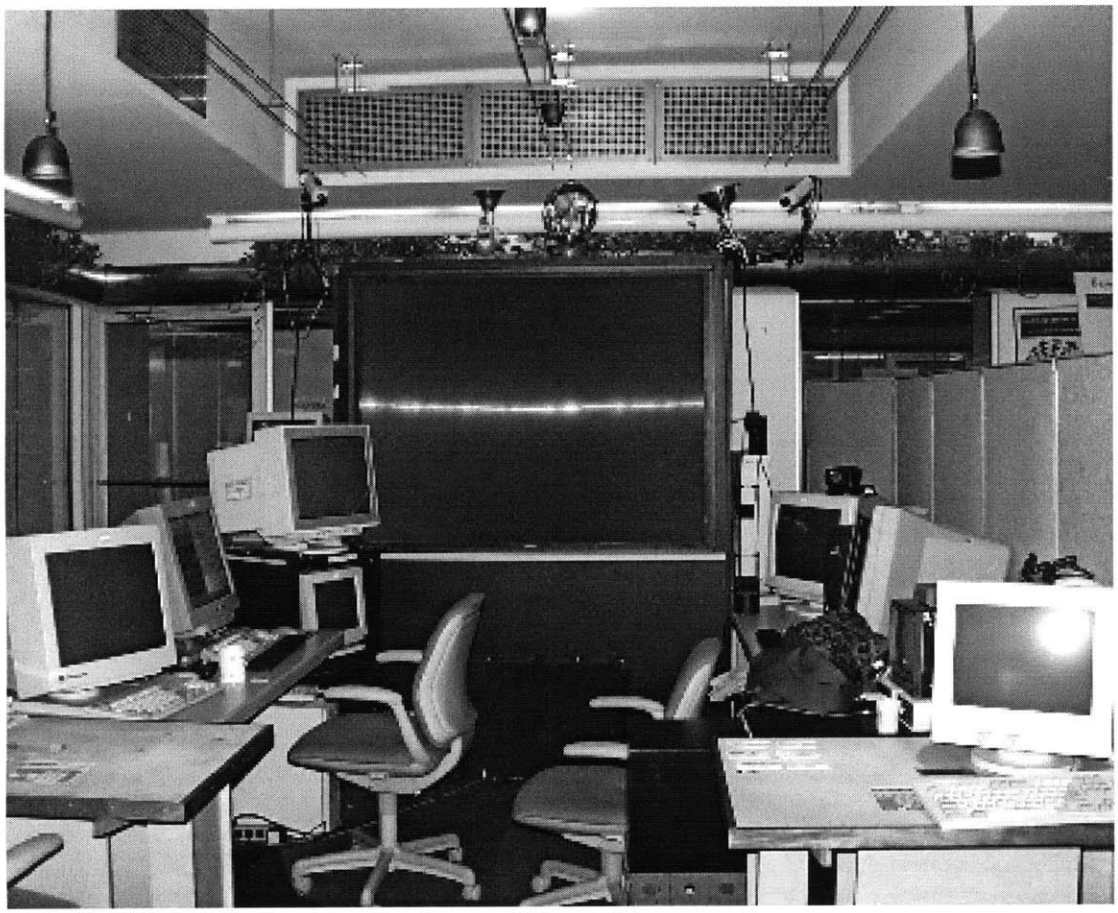

Figure 5-1: The REA system, showing the large screen display with two color cameras on top, and some of the workstations. The user interacts with REA by facing the screen and talking to her image. The user's voice and intonation are currently picked up by a lapel-worn microphone, and gesture data is collected by the stereo cameras.

video from the cameras, one dual processor SGI Octane workstation for sentence planning and general computation, a Pentium PC running IBM Via Voice for speech to text, a Pentium PC running Festival for text to speech, a Pentium PC for graphics, a Pentium PC for discourse planning, a stereo audio amplifier, and a pair of loudspeakers.

REA is designed to accept input in two main modalities: speech and gesture (in the future we hope to add intonation, facing direction and gaze tracking). REA's output is a computer generated voice and image, currently generated by Festival and Open Inventor respectively, designed to control speech, intonation, lip shape, gaze, and gesture.

The gesture input is derived from the STIVE system [3], which tracks head and hands in three dimensions in realtime at about 10 to 20Hz via calibrated stereo video cameras [1]. The STIVE 3D data is fed through feature extractors, and these features go to a Hidden Markov Model (HMM) recognizer to classify gestures. In addition, there is a simple system which signals REA when (1) the user enters and departs the sensing area; (2) the user's hands go into gesture space or are in motion; and (3) when the users hands return to rest. There is another simple system for detecting "big deictics," described in Section 6.11.1.

REA's output is designed with the appropriate knobs for conversational interaction. She can nod and turn her head and move her lips, and make hand gestures as well. Festival has been modified to provide accurate phoneme timing, which we use to synchronize speech and gesture.

REA's internal processing is a modular system [11] with some modules devoted to managing the interactional components of conversation; ands other modules for representing the topic component of conversation, information about the user's state (e.g. what the user has already seen and heard about), and for planning what steps are necessary to achieve REA's goal of conveying information describing the house.

[1] one of the features of STIVE is a relatively painless calibration process

```
time      head position          left hand position     right hand position

81094.5  -15.269  7.721  1.285   -1.581  6.524  1.476   -7.834 13.712 13.442

81169.5  -15.261  7.750  1.079   -1.738  6.601  1.298   -7.869 13.741 13.458

81173.2  -15.266  7.729  1.102   -1.752  6.693  1.298   -7.863 13.756 13.469
```

Figure 5-2: Three lines of raw stive data.

5.2 Offline Training data collection, and training of HMMs

As described in section 4.4, gesture data was gathered from naive subjects whose hand motions were tracked as they were interviewed about various places where they had lived or worked. The subjects were placed in front of a video display where they saw the interviewer's face. There was also a STIVE system running which tracked their head and hands in 3D and recorded the data (plus a timestamp) in raw tracking data files for later analysis. A video recording was also made using a quadsplitter to record reduced version of three video scenes plus an audio track. The three scenes were: both STIVE camera views of the subject, and a graphical representation of the tracking data consisting of three ellipsoids representing the head and hands, plus an elapsed time clock corresponding to the timestamps on the raw tracking data files.

STIVE represents each tracked object as an ellipsoid in 3D space. Figure 5.2 shows three rows of data from the raw data file labeled `debc2` (additional carriage returns and spaces have been inserted to make the data more readable):

Each row has a timestamp, in milliseconds from the when the tracking program started, followed by three groups of three numbers, which are the x,y,z positions of the head and hands, measured in the coordinate system of figure 5-3 and expressed in inches. Notice that the difference between the first two timestamps is about 75 msec, while the difference between the second two is about 4 msec. One of these gaps is several video frames; the other is less than a frame. Multi-frame gaps occur when STIVE has to do extra search to locate an object between frames, often because the object has moved farther than predicted by STIVE's inter-

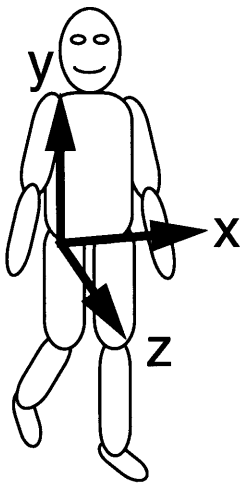

Figure 5-3: The coordinate system used by the STIVE tracking software.

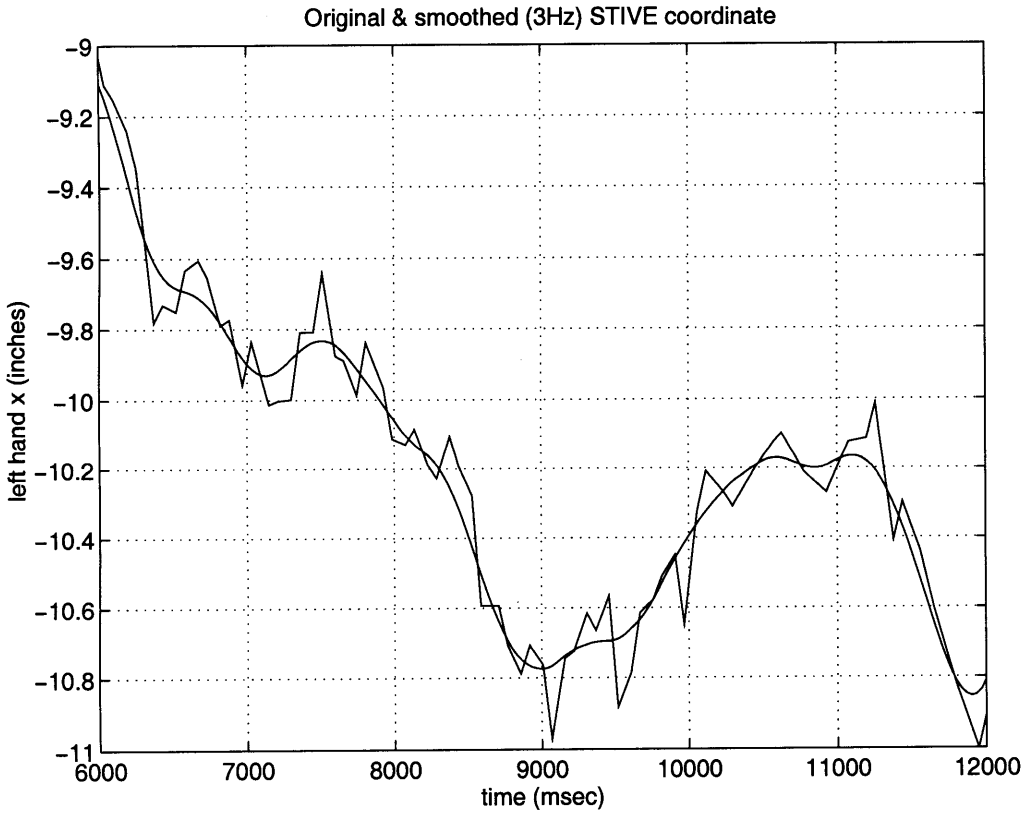

Figure 5-4: Six seconds of STIVE data, before and after resampling and low-pass filtering.

nal Kalman filter. Sub-frame gaps occur because there is jitter in STIVE's timestamp. These problems, plus noise on the position measurements, are dealt with by resampling the data to a uniform time base, and then filtering the positions through a lowpass filter. Figure 5.2 shows some STIVE data, before and after filtering.

Once the raw data was collected, I annotated all the gestures that STIVE tracked in the file, by viewing the video, listening to the audio, and marking the segments in the tracking file during which gestures happened. This took on the order of 1 hour of annotation time per minute of recorded data. Some of the annotations were audited by my advisor. The annotations were collected in annotation files corresponding to the raw tracking data files. Here is a small part of the annotation file covering the raw data presented above:

```
.1600           glitch
74.8848         rest

75.8170         diconic         "you came in the door"
77.0748         rest

77.7617         diconic         go in the door
78.4148

81.0945         rdeict          "you went down a flight of stairs"
82.2173
```

The first column is the timestamp in seconds, the second column is the gesture name, and the rest of the data on the line is a comment, usually indicating what the subject was saying during the gesture (this can be helpful as a sanity check, both for keeping track of one's place in the file, and when double checking annotations). During the first minute of this conversation, the subject had her hands positioned where STIVE could not track them, so that interval is annotated as a large glitch. Once the subject began talking and became animated, she made two double-handed iconic "`diconic`" gestures, followed by a right-handed deictic.

Some periods are annotated as "`rest`," while others have no annotation at all. If an annotation is blank, the corresponding time period is treated as a rest. Thus every instant of the collected data has some kind of annotation, either rest or glitch, or a gesture. The annotated

gestures are: beat, deictic, iconic, metaphoric, butterworth, preparation, and retraction. Each of those can occur either on the left hand, the right hand, or both.

5.2.1 Preparing the Data for Training

The raw data is not suitable for training HMMs: it is sampled at irregular intervals and it is noisy. These two problems are dealt with by (1) computing a cubic spline that passes through all data samples and then resampling the spline at every sixtieth of a second; and (2) filtering all the position measurements through a 3Hz lowpass filter. The filter is chosen to have an impulse response with very low ripple, as opposed to a filter with a sharp frequency cutoff. Filters with sharper cutoffs tend to have more ripple, and this shows up as ringing around transients. When derivatives are taken, if the signal has ringing, the derivative operation will amplify the ringing, resulting in a noisy velocity signal. A low ripple filter produces a cleaner velocity signal.

The annotation file is then used to chop the data stream into separate files where each file contains the smoothed evenly sampled XYZ data for an individual labeled gesture. These gesture files are named with a subject code and a sequence number, to help trace errors back to the raw data and annotations. For example, the `rdeict` gesture above was written into the file `debc2r_006`.

The individual data files are then processed to extract different feature sets, which are written into data files readable by Entropic Software's HTK suite of HMM tools. For example, processing the `debc2r_006` file generates a set of HTK training files such as `debc2r_006_cdthz.R.ext`, `debc2r_006_abspolar.R.ext`, and `debc2r_006_delta.R.ext`; where the names now also indicate the feature and the hand that made the gesture. For a two handed gesture such as the `diconic` above, twice as many files would be written - for each feature set and for each hand. Features, as in [9] were calculated as follows:

Polar coordinates - the y (vertical) coordinates of the hands were left unchanged, but the x and z coordinates were used to calculate the radius from a pole defined by a vertical line through the head, and the angle between a ray from the pole to the hand and a ray from the pole in the +x direction.

Velocities - computed numerically from the resampled points using symmetric first differences $(1/2(n(t+1) - n(t-1)))$.

Autocorrelation - let us call the most recent approximately 1 second segment of data for a particular feature the probe vector. The probe vector is correlated with other vectors of data from the same feature that start earlier in time over a range of about two seconds by computing a normalized dot product between the probe vector and the other vector. Since the probe vector will correlate perfectly with itself, and often very highly with vectors with small timeshifts from itself, the correlations are weighted by their timeshift thus: if N correlations are computed, the one timeshifted by N points has weight 1.0, while the one with zero timeshift has weight $1/N$. The autocorrelation operation returns two values, the weighted correlation with the highest absolute value, and its corresponding timeshift.

5.2.2 Training the HMMs

As explained in section 4.3.1, the size of the training data set determines the length of the feature vector (in short, too long a feature vector leads to overtrained HMMs, which can be detected by comparing the cross validation results versus the results from training and testing on the whole dataset). As mentioned in section 4.3.2, the approach of treating each hand separately may allow more features per hand for a given training set. For the data gathered for this project, separating the hands proved better than combining them, and vector lengths of 3 and 4 were tested for overtraining, resulting in a selection of feature length 3.

Table 5.1 shows how many training samples there are for each of the 8 categories of gestures. Note that with only 8 examples of butterworths there are too few to reliably train an HMM.

HMMs were trained using Entropic Software's HTK suite of HMM tools. Different feature vectors were compared via cross validation by quarters. In other words, the data set was divided into quarters, and HMMs were trained on three quarters and tested on the remaining quarter. This was done a total of four times, each with a different quarter in the test set, so that ultimately all the data was tried in the test set, and results were accumulated over all four trials. The advantage of cross validation is that it allows all the data to participate in the test set, resulting in the most thorough test of the data set.

Gesture category	Total count
rest	121
beat	132
preparation	51
retract	49
metaphoric	66
butterworth	8
deictic	34
iconic	209
total	670

Table 5.1: Counts of the different kinds of gestures in the training set according to hand annotation. Note: as explained in the results section, rests are far more common than any other category, but their number in the training set was limited to prevent them from swamping the other results.

The HMM features that scored best in cross validation tests are used by REA for gesture recognition; however REA's HMMs were trained on all of the data.

As noted above, the methodology could not distinguish between metaphoric and iconic gestures, so although separate HMMs were trained for these two categories, their outputs are merged into the category of "illustrative" gesture.

5.3 Overview of STIVE

STIVE stands for Stereo Interactive Video Environment. It is a system for tracking the positions of the hands and head of a single user in a volume of space imaged by a pair of video cameras, in 3 dimensions, in real time. STIVE models the user as a set of three ellipsoids, left hand, right hand, and head, each of which is characterized by a set of 10 parameters consisting of and x, y, z centroid, three shape parameters that measure the axes of the ellipsoid, and four quaternion parameters that indicate the orientation of the ellipsoid.

When STIVE is queried it returns a timestamp and a measurement consisting of three sets

of 10 parameters. The rate at which it returns data is variable – when it is tracking well it returns 10 to 20 measurements per second or more; when it has lost the user and needs to search widely it can slow down to less than one measurement per second.

STIVE consists of the following programs: ffinder, sffinder, and, optionally, bodymodel. Here is what each program does:

- **ffinder** ffinder, also known as flesh finder, takes input from a video camera and searches the current video frame for "flesh colored" pixels. flesh color is learned from training images using the coloredit program, and is defined by a gaussian distribution in the two dimensional colorspace $(u/y, v/y)$ where y is luminance (black and white), and u and v are color differences. [2] The purpose of representing flesh color this way is to provide some independence to variations in light intensity and melanin density.

  ffinder uses connected components analysis to group the flesh colored pixels into three elliptical "blobs" presumed to be two hands and a face. All this pixel processing is compute intensive; furthermore two ffinders must run – one for each camera – hence each ffinder runs on its own SGI O2 computer. To reduce search and speed up computation, ffinder uses blob position and velocity from the previous video frame to pick a search area to look for the blob in the next frame. If the blob is not found this way, ffinder must revert to a whole image search which is much slower and can produce only a few measurements per second. If less than three blobs are found, such as when one hand is hidden, or a hand becomes contiguous with the face or other hand in an image, then ffinder reports no measurements. This behavior is modified if the optional bodymodel module is running.

  When ffinder is running nominally and finding three blobs it reports five parameters for each blob to sffinder. The five parameters are: (x, y) centroid, ellipse major and minor axes, and angle of ellipse major axis. ffinder also provides a labeling guess of which blobs are left, right, and head.

- **sffinder** sffinder, also known as stereo ffinder, takes blob measurements and labels from

[2] The color transformation from RGB to YUV is done in hardware on a Silicon Graphics O2 according to these equations: $y = 0.257r + 0.504g + 0.098b + 16.0$, $u = -0.148r - 0.291g + 0.439b + 128.0$, $v = 0.439r - 0.368g - 0.072 * b + 128.0$

the two ffinders, and computes the ten 3D parameters for each blob (position, shape, orientation). In practice, the 3D shape measurements are very noisy, so they have been set to constant values. In order to compute 3D parameters, sffinder must have camera calibration information about how the two cameras are positioned and pointed, and their field of view in the STIVE sensing space, relative to some coordinate system. The camera calibration is an offline process in which a special version of sffinder records a person moving about in the sensing space and estimates calibration parameters consistent with the recorded motion. Calibration information only needs to be recomputed when the cameras are moved, and it is used every time STIVE runs. sffinder is not very compute intensive, so when both ffinders are running nominally it provides the 30 STIVE measurements plus timestamp at a rate of 10 to 20 or more measurements per second. If either of the ffinders cannot see enough blobs, sffinder reports "person not present" instead of measurements. This behavior is modified if the optional bodymodel module is running.

- **bodymodel**, is an optional module that brings knowledge of 3D human body dynamics to STIVE. This improves STIVE's performance in two main ways. First, bodymodel provides better predictions to the ffinders about where to search for the blobs. ffinder used a predictor that operates in the image plane, while bodymodel predicts in 3D using a model of the human body, and then projects search hints down into the image plane. Second, bodymodel handles some cases where two blobs become contiguous, and thus allows ffinder augmented by bodymodel to keep tracking in some cases where ffinder alone would fail. Both of these enhancements allow STIVE to do, on average, less search, and thus provide more frequent measurements.

5.4 HMM Recognition

The HMM recognition program is called stivetorea. It calls STIVE using a remote procedure call (RPC) connection, and receives the 30 STIVE measurements plus timestamp. It processes the hand coordinates to extract features, accumulates strings of features over time, runs the features through Hidden Markov Models to classify gestures, and and passes the classifications on to REA using KQML, the Knowledge Query and Manipulation Language [22].

5.4.1 Resampling and filtering

The measurements obtained from STIVE are accumulated in blocks of at least 0.4 seconds, then resampled in time using a cubic spline fit, then lowpass filtered with a 23 tap FIR filter. The length of the filter governs the choice of how many STIVE measurements to accumulate before resampling and lowpass filtering – 23 taps at 60Hz is just under .4 seconds. This essentially duplicates the filter used in Figure 5.2, which is used for the offline training data.

Chapter 6

Results

In general, the HMM recognition results of this project were disappointing. All of the cross validation tests achieved a recognition rate below 60%. Random chance would achieve a recognition rate of only 20%, so this result does prove the conjecture that there is significant independent information conveyed by observation of spontaneous co-verbal gestures. However, the recognition rate is not high enough to be useful as a standalone input channel.

On the other hand, the modifications of REA to respond to gesture and speech work well when the recognition system succeeds. REA responds to the user's turntaking cues, deictics, and certain beats.

Most of the HMM experiments were done treating the two hands independently to increase the training set and reduce the number and complexity of models, as described in section 4.3.2. Treated in this manner, there are eight gesture categories, and 670 gestures.

As tabulated above, the breakdown of gestures is:

rest	beat	prepare	retract	deictic	iconic	meta	butterworth
121	132	51	49	34	209	66	8

A note about the rests: the subjects all spent the largest proportion of their time with their hands resting. Some rests lasted 20 or 30 seconds. However, if all these rests were included in the training data, they would skew the results - a system that always said "rest" might be correct 80% of the time, but it would hardly be useful for gesture classification! On the other hand, eliminating rests from the test sample would obscure important rest / beat confusions.

Therefore, a subset of rests was chosen - those with lengths between 0.33 and 2.0 seconds. This range of lengths was chosen because it selects rests of lengths comparable to the other gestures, and will thus help identify confusions. Furthermore the choice yields a number of rests comparable to the number of beats, and thus will not favor a bias towards either rests or beats.

Among the gestures, there were only eight butterworths, all from the same individual. When trying to train a butterworth HMM, this small training set often resulted in bad condition numbers for matrix inversion, indicating too little data to estimate a covariance matrix somewhere in the model. Therefore, butterworths had to be eliminated from the final results. As noted above in section 3.5, separate models were trained for iconic and metaphoric, but their results were merged in the final confusion matrices.

rest	beat	prepare	retract	deictic	iconic + meta
121	132	51	49	34	275

6.1 Offline HMM cross validation

All final results presented are cross validation tests where the gesture data was divided into quarters by assigning gesture n to data segment $nmod4$; then training on three quarters of the data while testing on the fourth, and repeating the train test cycle four times so that all of the data was tested; and then combining all the test results. This means that all of the gestures appear in the test set, but the training set and test set are always disjoint.

Cross validation is a defense against overtraining. Overtraining means that there is little training data compared to the freedom in the model, such that that the model adapts not to generalities of the training data but rather to peculiarities in the particular training set.

Comparing cross validation results with results from an experiment that trains on all and tests on all gestures is a way to check for overtraining. If the cross validation results differ significantly from the train-all test-all results, this is an indicator that the training set is inadequate.

In these experiments, where the two hands were treated entirely independently in order to enlarge the training set, with three element feature vectors and full covariance matrices,

the cross validation comparison indicates that the training set is adequate. The error rate for cross validation was 53% to 56%; the corresponding train-all test-all error rates were about 4% higher (56% to 60%).

The left column of figure 6-1 shows a typical confusion matrix for models trained and tested on all energy-filtered gestures using the dpolar feature vector. The right column shows the corresponding cross validation result.

When the hands were treated separately, with six element feature vectors and diagonal covariance matrices, the cross validation comparison indicated inadequate data: the best cross validation recognition rate was 43.5% (35% without merging similar semantic categories); the corresponding train-all test-all result was 61% (54% without merging). The large gap between cross validation and train-all test-all signifies inadequate training data for the complexity of the models.

6.2 Energy filtering

Numerous confusion matrices from all feature sets showed rest and beat gestures were frequently confused. Furthermore, in many of these confusions, the second choice was the right choice, and the second choice was close to the first by a small margin in log likelihood. The "energy" of the beat gesture versus the energy of noise was investigated to account for this confusion. Energy is proportional to the square of velocity, so average energy of a gesture is just the sum of the squares of the velocities in x, y, z scaled by the duration of the gesture.

As shown in figure 6-2, rest and beat have a large overlap in energy. Rest energy varies because the system noise changes depending on how well STIVE can lock onto the patches of flesh it searches for. Beat energy varies because some people made very tiny, brief beats at times when the system noise was low.

This large overlap in the low energy region gestures causes beats and rests to be frequently confused. In order to give the system a better chance of differentiating between rests and the gestures that involve actual movement, I filtered gestures based on energy. The most energetic 10% of rests were filtered out, and the least energetic 10% of the other gestures were filtered out This removed 12 rests and 55 low energy active gestures, of which 46 were beats, all from

Fdpolar, train and test on all gestures

Confusion mat: NBest for N = 1, thresh = 10.00
WORD: Corr=57.96, [H=346, D=0, S=251, I=0, N=597]

	r e s t	b e a t	p r e p	r e t r a	d e i c t	i c o n i	
rest	71	27	0	0	0	0	[72.4/4.5]
beat	20	62	2	4	4	16	[57.4/7.7]
prep	1	8	19	2	7	12	[38.8/5.0]
retra	6	10	2	23	2	4	[48.9/4.0]
deict	1	1	4	0	22	5	[66.7/1.8]
iconi	9	50	20	13	21	149	[56.9/18.9]

Confusion mat: NBest for N = 2, thresh = 10.00
WORD: Corr=72.53, [H=433, D=0, S=164, I=0, N=597]

	r e s t	b e a t	p r e p	r e t r a	d e i c t	i c o n i	
rest	94	4	0	0	0	0	[95.9/0.7]
beat	6	85	2	2	4	9	[78.7/3.9]
prep	1	6	30	2	1	9	[61.2/3.2]
retra	3	6	2	31	2	3	[66.0/2.7]
deict	1	0	2	0	27	3	[81.8/1.0]
iconi	9	49	15	11	12	166	[63.4/16.1]

Fdpolar, cross validation

Confusion mat: NBest for N = 1, thresh = 10.00
WORD: Corr=54.62, [H=325, D=0, S=270, I=0, N=595]

	r e s t	b e a t	p r e p	r e t r a	d e i c t	i c o n i	
rest	108	1	0	0	0	0	[99.1/0.2]
beat	24	33	4	3	5	16	[38.8/8.7]
prep	4	10	13	3	8	13	[25.5/6.4]
retra	7	6	2	19	2	9	[42.2/4.4]
deict	0	4	3	2	8	17	[23.5/4.4]
iconi	25	43	21	13	25	144	[53.1/21.3]

Confusion mat: NBest for N = 2, thresh = 10.00
WORD: Corr=67.90, [H=404, D=0, S=191, I=0, N=595]

	r e s t	b e a t	p r e p	r e t r a	d e i c t	i c o n i	
rest	109	0	0	0	0	0	[100.0/0.0]
beat	5	66	3	1	3	7	[77.6/3.2]
prep	4	5	22	3	6	11	[43.1/4.9]
retra	6	4	2	25	2	6	[55.6/3.4]
deict	0	3	0	2	15	14	[44.1/3.2]
iconi	21	35	19	8	21	167	[61.6/17.5]

Figure 6-1: Confusion matrices for dpolar feature vector, energy filtered dataset. the left column is results from models trained and tested on all gestures; the right column is crossvalidation results. The upper confusion matrix reports only the correct recognition results; the lower matrix reports if either of the two highest likelihood results was correct, as long as the log likelihood is within 10.0 (this is the "`thresh`" parameter). Within each confusion matrix, correct identifications appear on the main diagonal, and idenfication errors appear in other elements of the row (for example, in the upper left confusion matrix, 62 beats were correctly identified, and 20 were erroneously identified as rests). The column to the right of each matrix, appearing as `[X/Y]` contains the percent of correct identifications for that row, followed by the contribution of that row to the total error.

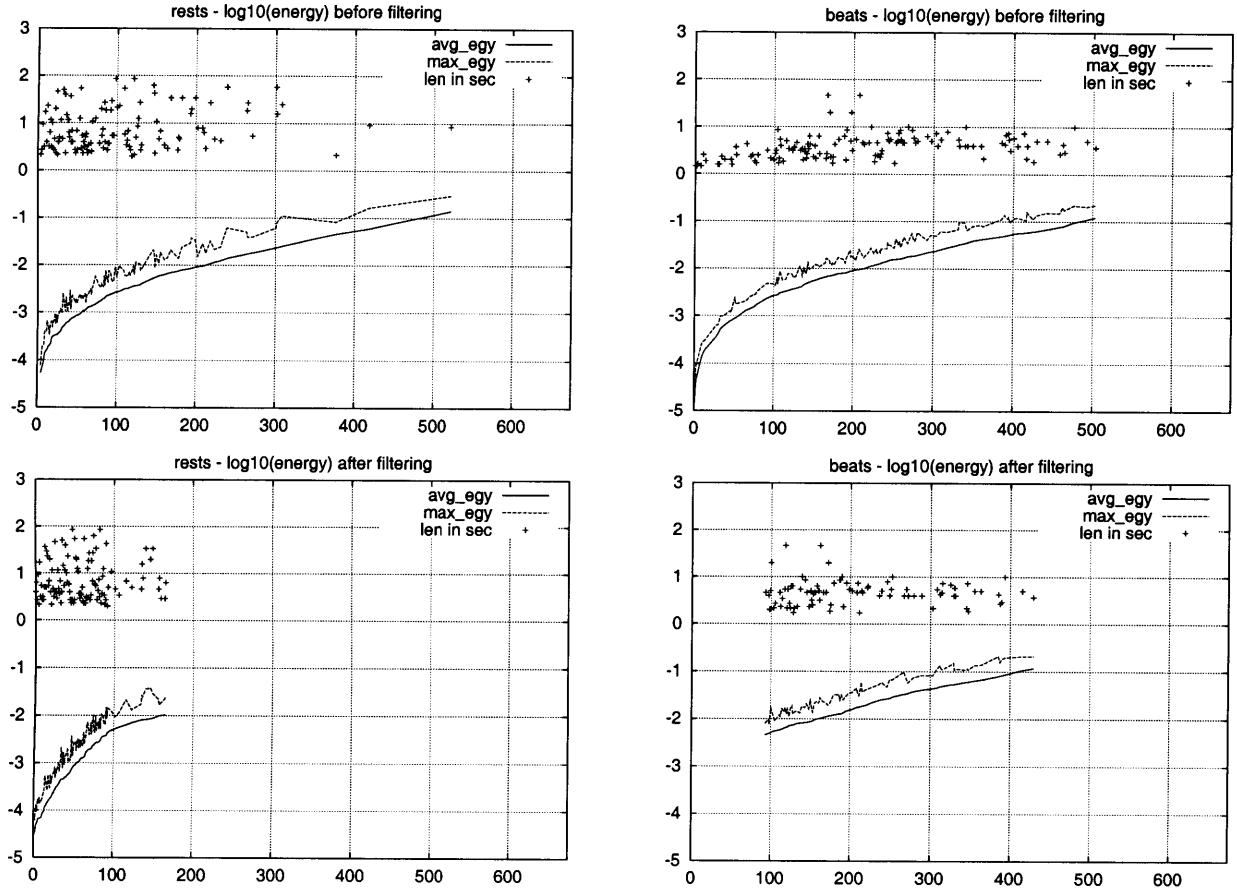

Figure 6-2: As can be seen by the upper two plots of log energy, rest and beat occupy similar energy bands. Each gesture in the category is sorted by log average velocity energy (solid line, arbitrary units). The dashed lines are peak energy, which would indicate gestures with major glitches (doesn't seem to be a problem). Finally the dots indicate how long a gesture lasts in seconds. The "density" of dots also indicates how many of the gestures are in that energy band.

a single subject (subject MH). With the eight butterworths also removed, this yielded a final training set of 595 gestures.

This produced a marked improvement in recognition rates. For example, before filtering, overall `dpolar` recognition was 42%, and the rest/beat confusion looked like:

```
rest   74   21      [61.2/6.7]
beat   32   65      [49.2/10.0]
```

After filtering, the overall recognition rate for `dpolar` was 54.6% and the rest/beat confusion looked like:

```
rest  108    1      [99.1/0.2]
beat   24   33      [38.8/8.7]
```

Note the asymmetry of errors: many more beats are mistakenly labeled rests than vice versa. A simple way to deal with this is by "boosting" the likelihood of beat so that the errors of beat and rest can be traded off. For example, with a boost of 0.8, the overall recognition rate drops slightly to 54.3% and the rest/beat confusion is:

```
rest  104    5      [95.4/0.8]
beat   18   41      [48.2/7.4]
```

Boosting beat by 2.0 yields an equal number of errors, with recognition rate of 53.45 and confusion:

```
rest   97   12      [89.0/2.0]
beat   12   48      [56.5/6.2]
```

Boosting beat by 2.5 yields a recognition rate of 52.94 and confusion:

```
rest   92   17      [84.4/2.9]
beat    9   53      [62.4/5.4]
```

Boosting beat by 3.5 yields nearly balanced error rates, with a recognition rate of 52.27 and confusion:

```
rest   85   24      [78.0/4.0]
beat    5   62      [72.9/3.9]
```

This demonstrates that with small changes in the likelihood, we can trade off rest versus beat errors, reaching a point where the errors are balanced, or where the percent correct is nearly balanced. The ability to trade off errors is useful when tuning REA's gesture recognition because some errors effectively have a higher cost (i.e. they are more "embarrassing" to REA), so trading off allows us to attempt to reduce cost for a given error rate.

After energy filtering, the revised breakdown [1] of gestures is:

	rest	beat	prepare	retract	deictic	iconic
old	121	132	51	49	34	275
new	109	86	51	45	34	270
change	-12	-46	0	-4	0	-5

6.3 HMM Re-Estimation with pre-segmented gestures

Re-estimation is the phase in the HMM training process in which the training data is reprocessed, and all HMM parameters (means, covariances, and transition probabilities) are updated. This phase is also known as embedded training, especially when the training data is not pre-segmented. When, as in this project, the training data is pre-segmented, re-estimation has less of an effect. As figure 6-3 indicates, 9 cycles of re-estimation changed the recognition rate by about 0.5%; a positive change in one case, and a negative change in another case. Because the effect of re-estimation is so small here, 9 cycles of re-estimation were deemed more than adequate for this project.

[1] This breakdown is for filtering on energy derived from Cartesian velocity $v_x^2 + v_y^2 + v_z^2$. For some experiments, energy was computed from different coordinates; hence there may be small differences in the gesture subtotals.

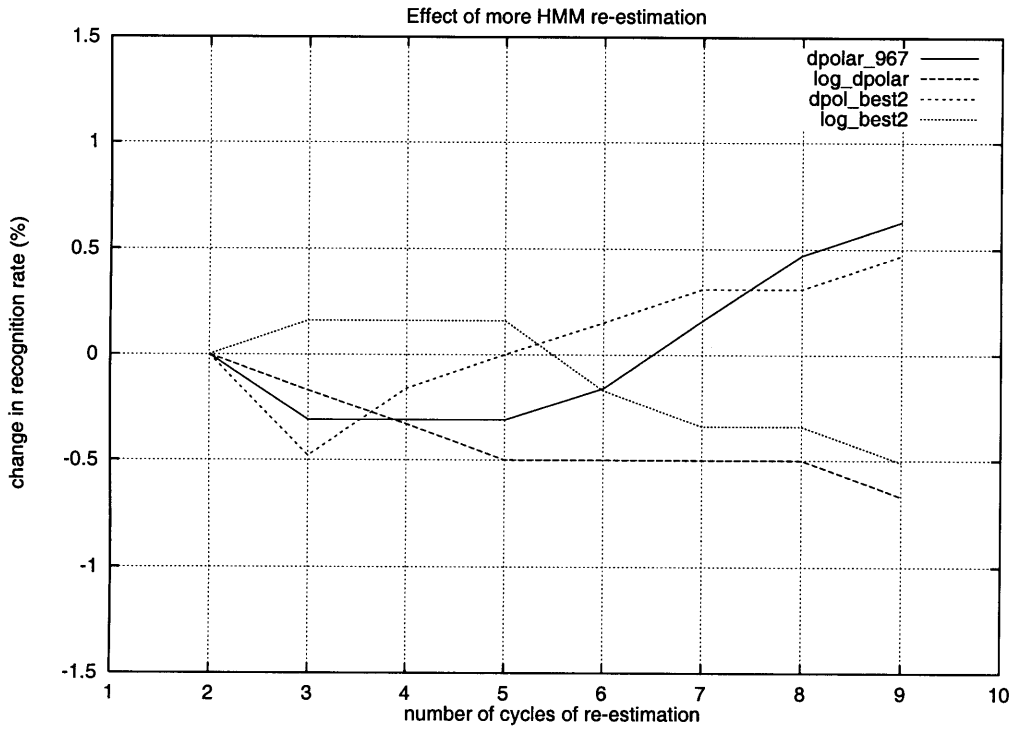

Figure 6-3: Re-estimation experiments. Cross validation recognition rates from different feature sets are recorded at each re-estimation step to see how it affects recognition rate, and changes in the rate are plotted. The four curves represent two feature sets (dpolar and log dpolar), with the best of 1 and best of 2 changes both being plotted.

6.4 Autocorrelation features

The results from the best autocorrelation run with energy filtering produced a recognition rate of about 53%, as shown below.

```
    xvalrun Fcdthz
-----------------------------------------------------
Confusion mat: NBest for N = 1, thresh = 10.00
WORD: Corr=53.04, [H=331, D=0, S=293, I=0, N=624]
-----------------------------------------------------
         r    b    p    r    d    i
         e    e    r    e    e    c
         s    a    e    t    i    o
         t    t    p    r    c    n
                        a    t    i
 rest   88   17    0    1    1    1  [81.5/3.2]
 beat   38   64    2    4    4    9  [52.9/9.1]
 prep    1   15   16    2    6    9  [32.7/5.3]
retra    7   11    4   19    2    5  [39.6/4.6]
deict    1    2    5    0    9   16  [27.3/3.8]
iconi    8   61   15   17   29  135  [50.9/20.8]
-----------------------------------------------------
```

The problem here is that when the feature vector is limited to three features, something must be displaced to make room for the autocorrelation feature. In this case, the correlation feature is how far back in time is the weighted peak correlation, where the correlation is weighted according to the scheme described in Section 5.2.1. It displaced radial velocity in a feature vector that was originally polar velocities (`dpolar`). Radial velocity was picked to be the least important feature in the polar velocity feature set. Although correlation may be a useful feature, it is clearly less useful than the various features I tried replacing it with.

6.5 Single user, single user excluded

One of the subjects from whom the raw training data was collected, MH, contributed a large minority of gestures, 284 after energy filtering. This is a large enough subset to do a cross validation run on this subject alone, as well as cross validation on the data from all subjects

excluding MH. This subject is a long-time, but not native, speaker of English, and a very fluid gesturer. Interestingly, he contributed all eight butterworths, and all but one of the 46 low energy beats excluded by energy filtering.

Training and testing on all of MH's gestures produced a recognition rate of 55.28%. Cross validation by quarters produced a recognition rate of 51.06% – close enough to the train-all test-all result to conclude that over training is not a significant issue.

Here is the confusion matrix for a cross validation test of MH only with energy filtered velocities as features:

```
   MH only, Filtered delta, cross validation
------------------------------------------------------------
Confusion mat: NBest for N = 1, thresh = 10.00
WORD: Corr=51.06, [H=145, D=0, S=139, I=0, N=284]
------------------------------------------------------------
         r    b    p    r    d    i
         e    e    r    e    e    c
         s    a    e    t    i    o
         t    t    p    r    c    n
                        a    t    i
 rest   51    5    0    1    5    3  [78.5/4.9]
 beat   12   34    2    1    3   11  [54.0/10.2]
 prep    0    4    6    0    7   12  [20.7/8.1]
retra    2    2    1    8    0    5  [44.4/3.5]
deict    0    2    1    0    2    8  [15.4/3.9]
iconi    7   11   10    6   18   44  [45.8/18.3]
------------------------------------------------------------
```

For comparison, here is the confusion matrix for a cross validation test of data from all subjects *except* MH, with energy filtered velocities as features:

```
   no MH, Filtered delta, cross validation
------------------------------------------------------------
Confusion mat: NBest for N = 1, thresh = 10.00
WORD: Corr=59.49, [H=185, D=0, S=126, I=0, N=311]
------------------------------------------------------------
         r    b    p    r    d    i
         e    e    r    e    e    c
         s    a    e    t    i    o
```

	t	t	p	r	c	n	
				a	t	i	
rest	44	0	0	0	0	0	[100.0/0.0]
beat	3	11	0	1	0	8	[47.8/3.9]
prep	2	4	8	1	2	5	[36.4/4.5]
retra	5	1	2	16	0	3	[59.3/3.5]
deict	0	4	2	0	6	9	[28.6/4.8]
iconi	17	28	9	11	9	100	[57.5/23.8]

One would expect that gestures from a single subject would be more homogeneous, and therefore yield better results than for all subjects combined, but this turned out to be false. Part of the explanation lies in the relative proportion of beats, and beat / rest confusion. Beats contribute 10.2% of the error in the MH only run, while they contribute only 3.9% of the error in the no MH run, despite having a worse overall error rate in the no MH run. This is because there are only 23 beats in the no MH data, while there are 63 beats in the MH data.

Preparation gestures mis-identified as iconics also contributed to the MH error. A possible explanation for this phenomenon is that MH's iconics may have more variety than all the other subjects combined. If they spread out more in the feature space, then they will be more likely to fall into other nearby categories, and may capture more gestures from the other categories. MH was sitting at a table during data collection, so there was a shorter distance from the resting position to his gesturing region. This shorter reach may have caused his prep and iconic categories to be more similar and more easily confused.

Finally, the no MH rests were identified without error. This could be because MH occasionally rested with his elbows on the table and his hands in gesture space, where they may have been moving slowly. This elbow rest position may have allowed him to make smaller (lower energy) gestures than the subjects lacking a table (the first three of seven subjects used a table, but the first two were eliminated because I was learning to annotate on their data; thus MH is the only subject in the final data who used a table). When people didn't have a table, they tended to rest their hands in their laps or on their knees, and were less likely to let their hands dwell in gesture space between gestures. These people were more likely to make gestures as single continuous movements, rather than a series that could be broken down into preparation, stroke, and retraction.

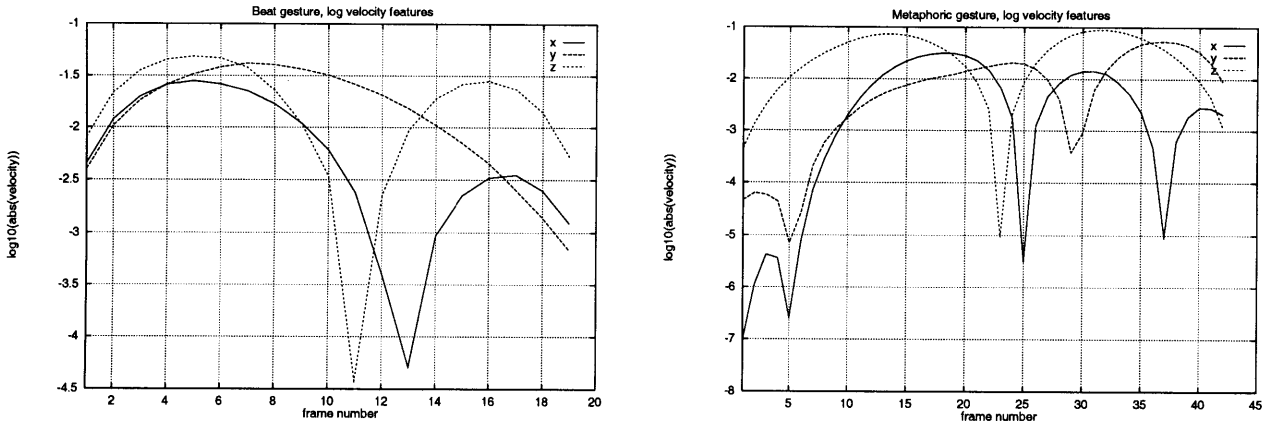

Figure 6-4: Plots of the log of the absolute value of velocity for a right hand beat and the left hand of a two handed metaphoric gesture.

6.6 Logarithm of velocity features

The log velocity feature set was an attempt to compute a feature with a more gaussian distribution of values. Figure 6-4 shows some sample raw gesture data of a beat and a metaphoric. Here is a sample confusion matrix:

```
   filtdeltalog10, cross validation
------------------------------------------------------------
Confusion mat: NBest for N = 1, thresh = 10.00
WORD: Corr=52.43, [H=313, D=0, S=284, I=0, N=597]
------------------------------------------------------------
         r     b     p     r     d     i
         e     e     r     e     e     c
         s     a     e     t     i     o
         t     t     p     r     c     n
                           a     t     i
 rest   88     9     0     1     0     0   [89.8/1.7]
 beat   12    52    10     9     4    21   [48.1/9.4]
 prep    0     5    11     9     8    16   [22.4/6.4]
retra    2    10     7     6     9    13   [12.8/6.9]
deict    0     2     6     2     6    17   [18.2/4.5]
iconi    0    27    21    31    33   150   [57.3/18.8]
------------------------------------------------------------
```

The log velocity feature was chosen because it may have a more gaussian distribution of values than pure velocity. Velocities go both positive and negative, so the absolute value must

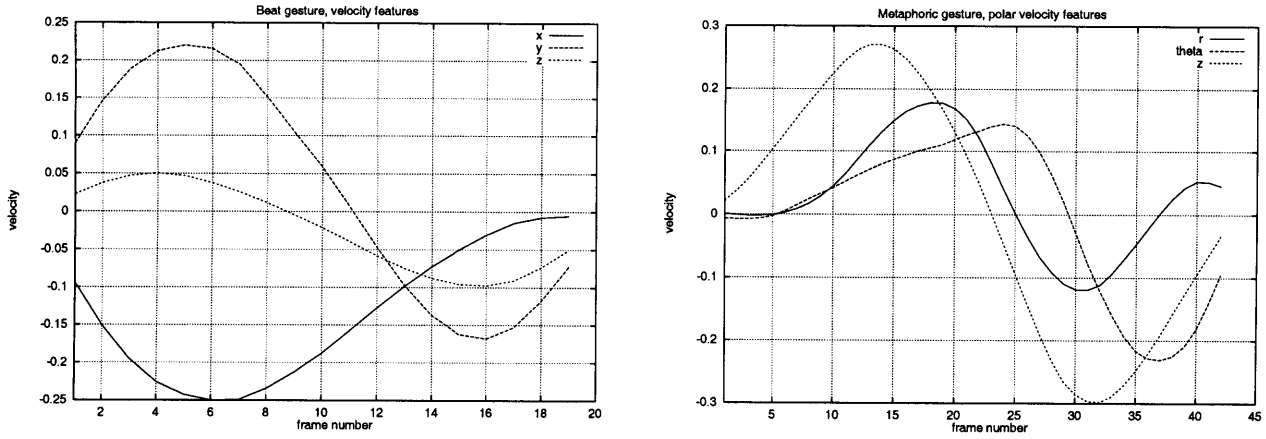

Figure 6-5: Plots of the velocity for a right hand beat and the left hand of a two handed metaphoric gesture.

be taken before the log. When the velocity has a zero crossing, it shows up on the log plot as a downward spike. These downward spikes are very salient features for the HMMs – if they are too salient, the HMM ends up training mostly on the number of zero crossings in the gesture, and is not influenced enough by other features of the gesture.

6.7 Best results

Figure 6-5 shows some sample raw gesture data of a beat and a metaphoric gesture for Cartesian velocity features. Figure 6-6 shows cross validation results for the two feature sets with best overall recognition rate.

Note that in both of them the sum of correct recognitions of rest + beat is a little over 140. Using the "boosting" technique described in section 6.2, errors in these two can be traded off, but the sum of correct responses stays in a range near 140.

The preparation and retraction gestures are included as separate categories only so they won't be confused with the gestures carrying meaning: beats, deictics, and iconics / metaphorics. Thus for purposes of understanding communicative intent, preparations and retractions can be merged with each other, and with rests. Doing so yields figure 6-7.

Filtered delta, cross validation

Confusion mat: NBest for N = 1, thresh = 10.00
WORD: Corr=54.29, [H=323, D=0, S=272, I=0, N=595]

	r e s t	b e a t	p r e p	r e t r a	d e i c t	i c o n i	
rest	104	5	0	0	0	0	[95.4/0.8]
beat	27	39	4	2	5	9	[45.3/7.9]
prep	4	9	20	1	7	10	[39.2/5.2]
retra	6	5	4	22	1	7	[48.9/3.9]
deict	2	4	4	1	9	14	[26.5/4.2]
iconi	22	49	20	24	26	129	[47.8/23.7]

Filtered dpolar, cross validation

Confusion mat: NBest for N = 1, thresh = 10.00
WORD: Corr=54.62, [H=325, D=0, S=270, I=0, N=595]

	r e s t	b e a t	p r e p	r e t r a	d e i c t	i c o n i	
rest	108	1	0	0	0	0	[99.1/0.2]
beat	24	33	4	3	5	16	[38.8/8.7]
prep	4	10	13	3	8	13	[25.5/6.4]
retra	7	6	2	19	2	9	[42.2/4.4]
deict	0	4	3	2	8	17	[23.5/4.4]
iconi	25	43	21	13	25	144	[53.1/21.3]

Figure 6-6: Confusion matrices for energy filtered dataset, delta and dpolar features.

Filtered delta, cross val., merged

Confusion mat: NBest for N = 1, thresh = 10.00
WORD: Corr=56.81, [H=338, D=0, S=257, I=0, N=595]

	r e s t	b e a t	d e i c t	i c o n i	
rest	161	19	8	17	[78.5/7.4]
beat	33	39	5	9	[45.3/7.9]
deict	7	4	9	14	[26.5/4.2]
iconi	66	49	26	129	[47.8/23.7]

Filtered dpoar, cross val., merged

Confusion mat: NBest for N = 1, thresh = 10.00
WORD: Corr=57.31, [H=341, D=0, S=254, I=0, N=595]

	r e s t	b e a t	d e i c t	i c o n i	
rest	156	17	10	22	[76.1/8.2]
beat	31	33	5	16	[38.8/8.7]
deict	5	4	8	17	[23.5/4.4]
iconi	59	43	25	144	[53.1/21.3]

Figure 6-7: Confusion matrices for communicative gestures, energy filtered dataset, delta and dpolar features, with preparations and retractions merged into the rest category. Since we treat preparations, retractions, and rests the same, confusions between these categories don't lead to system errors. Merging the categories provides a better measure of when gesture identification errors lead the system to behave erroneously.

6.8 Train on four subjects, test on fifth

To estimate how the system would work on subjects not present in the training data, a version of cross validation was run that trained on four subjects and tested on the fifth. The results of this experiment are:

```
   Fdpolar, train on 4, test 5th
-----------------------------------------------------------
Confusion mat: NBest for N = 1, thresh = 10.00
WORD: Corr=47.91, [H=286, D=0, S=311, I=0, N=597]
-----------------------------------------------------------
         r    b    p    r    d    i
         e    e    r    e    e    c
         s    a    e    t    i    o
         t    t    p    r    c    n
                        a    t    i
 rest   91    3    0    4    0    0  [92.9/1.2]
 beat   62   13    5   10    5   13  [12.0/15.9]
 prep    4   13   15    1    5   11  [30.6/5.7]
retra    8   11    2   14    3    9  [29.8/5.5]
deict    0    5    8    1    6   13  [18.2/4.5]
iconi   20   34   20   17   24  147  [56.1/19.3]
-----------------------------------------------------------
```

This is somewhat worse than the other cross validation results. The results from individual subjects indicate MH is an outlier subject:

Subject	DC	DS	MH	RS	SB
Recognition rate	55.43	66.15	40.54	49.53	51.35

Subject MH contributed nearly half the gestures, and the 40.5% recognition rate on his gestures dragged down the rest of the data. I re-ran the leave one subject out data omitting subject MH and got the results below: much more comparable to the 59.5% (see Section 6.5) crossvalidation results obtained by omitting MH.

```
   Fdpolar, leave 1 out, no MH
-----------------------------------------------------------
Confusion mat: NBest for N = 1, thresh = 10.00
WORD: Corr=55.15, [H=166, D=0, S=135, I=0, N=301]
```

	r e s t	b e a t	p r e p	r e t r a c t i o n	d e i c t	i c o n i	
rest	33	3	0	0	0	0	[91.7/1.0]
beat	0	6	2	5	3	7	[26.1/5.6]
prep	1	8	5	0	0	8	[22.7/5.6]
retra	3	6	2	11	2	4	[39.3/5.6]
deict	0	1	4	1	5	9	[25.0/5.0]
iconi	6	22	8	12	18	106	[61.6/21.9]

6.9 Applications of Gesture Classification

Classification of gestures via HMMs is one component of this thesis; the other component is the use of labeled gestures to aid REA in understanding the user's communicative intent. The following sections describe how REA interprets and responds when she detects combinations of gesture and speech.

6.10 Detecting and Using Interactive Information

In the present implementation, the interactive information REA recovers from hand gestures is turntaking. There are three turntaking cues that the REA system currently recovers from hand gestures:

- User has turn, user stops speaking, but hands moving or in gesture space indicates the user wants to hold the turn – REA should keep listening.
- User has turn, user's hands dropped, combined with end of speech markers, means user wants to give up turn – REA can talk.
- REA has turn, user's hands moving or in gesture space indicates the user wants to take or hold the turn – REA should finish sentence and give up turn.

There is an additional turntaking behavior in which REA stops speaking whenever the microphone detects the user speaking. The sum of all these behaviors makes REA appear to be very polite because she so freely yields the turn.

6.11 Detecting and Using Content Information

Content information is information about the topic of discussion, rather than information about the interaction. In general, the majority of content information is carried verbally, but there is some information that is carried gesturally. The gestural channels of content information are classified into three categories: deictics (pointing), emphasis beats, and illustrative. The problem of understanding illustrative gestures is beyond the scope of this thesis.

6.11.1 Detecting and Using Deictics

We created a special class of deictics, known as "large deictics" to handle the case when a user points to an object on REA's display screen. These deictics are recognized by a special thresholding and hysteresis procedure, rather than by HMMs. The reason HMMs were not used is that we collected no straight arm pointing gestures in the training set. All the deictics exhibited by the subjects were small gestures pointing at small imaginary objects, such as a chair or window in an imaginary room in the air in front of the subject. All these deictics were directed as small "models" the subject was describing.

We defin large deictics to be straight arm gestures directed at objects on REA's large display screen. To detect them the system first estimate the user's arm length, using the STIVE measurement of head height and scaling standard body measurements obtained from a table of human body measurements [18]. The system begin tracking a deictic when ever the user's hand moves beyond 1/2 arm's length. The endpoint of the deictic is defined as the maximum extension the hand reaches in the 1/3 of a second after the hand breaks the 1/2 arm's length cylinder. The direction of the deictic is defined as the user's estimated shoulder position, which is obtained by assuming the user's shoulders are parallel to REA's screen, and using the human body measurements to approximate the displacement of the shoulder from STIVE's measurement of the head position. In order to prevent a flood of deictic messages

when the hand is near the 1/2 arm's length boundary, hysteresis is introduced by requiring the hand to return within .4 arm's length before starting another deictic.

The user, exhibiting a deictic, points to an object on the large display screen and asks "what is that?" REA's gesture subsystem, detecting a deictic, computes a directed line along the user's arm. REA's understanding module attempts to resolve a referent for the demonstrative 'that.' Finding a deictic at the same time as the demonstrative, the understanding module passes the directed line to the graphics and animation module, which intersects the vector with REA's scene, computes the first collision, and returns the label of the object intersected by the vector (or null if the object is unlabeled). If the understanding module obtains a valid label, it will cause an explanatory sentence to be created and expressed. For example: in response to various deictics, REA says "That is a new kitchen island," or "The sofa is not for sale!"

6.11.2 Uses of Beats to convey communicative intent

emphasis

Attending to emphasis cues can allow the system to make guesses about the intent of the user, and thereby reply more naturally. Emphasis can be conveyed by intonation or gesture or both. In this example, suppose the speech to text system recovers the question:

"must the walls be blue and white?"

REA can reply to this with something like:

"We have many designs available."

But if emphasis information is available, the system can generate more appropriate replies:

"blue and white?" *"We have a lovely yellow and white ..."*

"blue and white?" *"We also have a blue and green ..."*

"blue and white?" *"You may choose a solid color ..."*

theme & rheme

Theme and Rheme (similar to the concept of "given and new") are a particular form of

emphasis which indicates which topics of a sentence the speaker considers to be part of the context or shared knowledge (theme), and which topics are newly introduced or spotlighted (rheme). For example, in the phrase "I walked through the door and saw a mouse;" the given information is 'the door,' a theme, part of the context of a house; while the new information is 'a mouse,' a rheme, the topic which the speaker wishes to spotlight.

rapport enhancing responses

Expressing areas of agreement can help to built rapport. REA makes use of beat recognition when she expresses agreement with certain preferences stated by the user. In this example involving kitchen detailing, the user may express a preference for either color or material or both, depending on which word (if any) is emphasized by a beat gesture. When the user says:

"I like blue tiles,"

If REA identifies 'blue' as the rheme, she'll respond:

"Blue is my favorite color."

On the other hand, if the user says:

"I like blue tiles,"

REA replies:

"I love tiles."

Finally, if there is no discernible emphasis when the user says "I like blue tiles," then REA responds "Me too!"

Note that the user says exactly the same sentence each time. REA demonstrates an enhanced understanding of the user's communicative intent by responding differently depending on which word (if any) is emphasized by the user with a beat gesture.

In summary the HMM recognition rate is disappointing because it is too low to be useful in a production system. On the other hand, REA's behavior in response to combined gesture and speech works as planned. REA provides appropriate responses to turntaking cues, deictics, and beats.

Chapter 7

Conclusion

As claimed, in this thesis project I developed an automatic system to classify, label, identify coincidence with speech, and respond to co-verbal gestures made by users in realtime while conversing with REA. While the HMM recognition rate is lower than desired, it is still far above the chance rate of 20%. This provides a measure of validity for the gesture categories used in the system. The system classifies the user's gestures without referring to the user's speech, identifies the word in the speech stream that coincides with the gesture, and then adds information from the gesture stream to the speech stream. Modifications made to REA allow REA to respond in a manner demonstrating an enhanced understanding of the user's communicative intent in cases involving turn-taking, deictics, and certain beats.

Thus, to the extent that a machine can recognize and use these categories of gestures to infer information not present in the words spoken, I have demonstrated that there is complementary information in the gesture stream and the relative timing of gestures with words.

7.1 What did the HMMs learn?

The main thing the HMMs learned concerns the relative duration of the gestures. Beats happen quickly, and the four state HMM and its transition probabilities reflect that fact. Iconics and metaphorics take longer, and the six state HMM reflects that.

Beyond that they learned something about the size of the gesture. The combination of

velocity features and lowpass filtering means that a brief gesture like a beat shows a small velocity, and a large iconic gesture shows a larger velocity.

Preparations and retractions have some consistent direction of velocity as well: preparations are mostly upwards, and retractions are mostly downwards.

7.2 HMM results were disappointing; here's why

Although the recognition rate results do prove the conjecture that there is significant information conveyed by observation of spontaneous co-verbal gesture, the overall best recognition rate results are disappointing, because recognition less than 60% is not good enough to be useful in a real world working system. Furthermore, the cross validation results for leave one subject out training and testing were even lower. Therefore, a system deployed in the real world would have an even more difficult task then the task in this thesis, so one would expect even worse results in the real world.

In a real world system, the training subjects will always be different from the test subjects. One can hope, as in speech recognition, to get a training set large enough to cover all the variability in the general population, but it is an expensive and grueling task – only worth undertaking if the expected payoff is large enough. And even in speech recognition, the system works better after being trained on the user's voice.

One of the problems encountered in this project was the confusion between beats and rest. It would be easy to blame this entirely on system noise, but not only do all systems have noise, many humans have twitches and involuntary movements, so noise cannot be eliminated even with some ideal tracking system. Furthermore, some beats can be very small, so one cannot expect beats and noise to be completely separable. Another problem encountered is that triphasic gestures - iconics and metaphorics - are sometimes ajacent to distinct preparation and retraction phases and sometimes not (i.e. sometimes one or both of those phases is elided, and sometimes one or both are combined with the iconic or metaphoric in a continuous motion). Perhaps there are regularities in behavior that would enable the system to predict when triphasics will be made in one continuous motion and when they will be broken up, but absent these regularities it is problematic to treat them as one class. Finally, tripha-

sics form a very diverse category. It would be helpful if they could be broken into visually separable subcategories, but that is a larger research project than this thesis.

Another issue is the gesture categories are not well enough defined. For example, here is the annotation of subject AM describing where she grew up:

```
101.4905        ldeictic        And then the living room.
103.0006

104.7601        rprep           and then
105.2400

106.9398        riconic         and then there's a hallway
107.5493

108.1297        dmeta           which sorta divides
109.2586

110.0289        rbeat           the house
110.4987

110.5288        rbeat           in half
111.2187
56.1695
```

The gesture for "in half" was annotated as a beat because it looked similar to other beats from this subject. But if she had performed one of her more vehement beats there, it could have been interpreted as iconic, lightly chopping the house in half. Another, less probable interpretation is that it was a deictic, pointing out the location of the hallway in some way. But the point is that the gesture associated with "in half" doesn't have to serve a single purpose. It can do all of the above! If most of its energy occurs during "half" it can be indicating that "half" is rhematic. It can simultaneously indicate a chop and a location.

In another example, subject DC describes an apartment and adds that it had a "river view," with a two handed beat on "river." But if the hand shape were a little more defined, she might have been iconifying the view. In fact, she may have been gesturing lazily, such that a lazy iconic looks like a two handed beat. Alternatively, she may be accomplishing both goals, iconifying and beating at the same time, giving the gesture the short time signature of

a beat, but also using the orientation of her arms to indicate that the good view was directly out the window, rather than e.g. far away and off to the side.

The straightforward assumption in gesture research is that the categories are mutually exclusive, and this assumption is built into REA's classifier. This assumption may not be correct.

7.3 Conclusion 2: A hard problem; more work is needed

The situation here is analogous to the early days of speech recognition. Vowels and consonants are a relatively small set of audibly different sounds, and the task of creating filters to identify them was deemed to be a not too difficult problem. Experience soon showed that was was "obvious" to the human ear was not easily captured by a filter. The first few years of progress in speech recognition were devoted to finding a set of features suitable for phoneme recognition. Only after that could HMMs be successfully applied to estimate word and sentence likelihoods.

The analogous situation for gesture recognition is that the feature sets tested here were not adequate, and an adequate set of features have not yet been identified. Unfortunately, the search for features requires a much larger training set than obtained here. It would be very helpful to be able to experiment with a longer (10 to 15 element) feature vector, because you could include multiple features that individually provide only a small discriminating ability. For reasons noted in section 4.3.1, scaling the size of the feature vector by n produces a need for n^2 times as much training data. Furthermore, since not all features have gaussian distributions, it would be useful to be able to handle non-gaussian features. A simple way to do that is to use a mixture of gaussians - the sum of several gaussians can approximate other distributions. However, a mixture of m gaussians has m times as many parameters and thus requires m times as much training data.

This thesis began with a list of characteristics of good features, including:

- Invariant to changes in viewpoint of camera
- Invariant to translation and rotation of user

- Able to model the "typical variation" in gesture performance
- Robust to "small variations" of gesture speed and shape

At this point we can add to the list. Characteristics of good features should also include:

- features that relate to hand shape;
- relative independence from the scale of a gesture;
- adaptivity to the style of the particular user.

Ideally one would also want continuous acquisition of and adaptation to a user model. This is a goal in speech recognition, and it should be in gesture recognition as well.

Another tool that would be helpful is discriminative HMMs – HMMs that train from negative as well as positive training examples. Commercial HMM packages available for this project only train on positive examples, and ignore the negative examples they misclassify. Discriminative HMMs are currently a hot research topic, and may be part of the "secret sauce" inside high end commercial speech recognition systems, but standalone DHMM tools are not yet available.

7.4 Observations on the gesture categories

The gesture categories used in this project – beats, deictics, iconics and metaphorics, butterworths, preparations, and retractions – seem to fall into two distinct groups.

In one group of gestures, each has a fairly well defined discourse function. Beats provide emphasis, deictics select items, and butterworths hold the floor while searching for a word. In addition, there are motor characteristics that give reason to believe that visual recognition of these gestures is feasible (though perhaps not sufficiently well recognized by a blob tracking system). Beats are small and brief movements; deictics usually involve a pointing finger, often parallel to the forearm; and butterworths involve repetitive motion, often at a higher frequency than a sequence of individual beats. Preparations and retractions may not have distinct discourse functions, but they have the well-defined mechanical function of getting

the hands into or out of gesture space. Preparations and retractions also have visual characteristics – preparations are usually upwards into gesture space, and retractions downward out of gesture space. Hence I will classify them in this first group with beats deictics and iconics.

Contrast the prior group of gestures with iconics and metaphorics. Iconics and metaphorics serve many and various discourse functions: they can provide descriptions like adjectives, for example a gesture indicating flat or smooth; they can provide action information, for example a typing gesture accompanying "I'll let you know;" they can provide path information; and, by drawing shapes in the air, they can provide complex spatial and shape information. Simultaneously the degree of vehemence can provide emphasis information. There are no particular visual characteristics of iconics and metaphorics except that they often longer than beats, not shaped like deictics, and not as repetitive as butterworths. The same shaped iconic or metaphoric can mean many different things in different contexts, and a listener must draw upon the context as well as the speech and gesture to understand many iconics and metaphorics.

Thus the second group is very different from the first group. In the first group gestures a human observer may be able to classify into their categories either by the gesture's function or purpose; or by the gestures visual characteristics. In the second group, the gestures cannot be classified without the associated speech, their shapes are more defined by what they are not, and there may be many domain specific sub-categories. For example, one will find path and shape gestures when someone is describing a route with landmarks; action pantomimes when someone is describing cartoon or a boxing match; and abstract emotive gestures when someone is describing a dramatic incident. The second group of gesture categories now appears to be a kind of catchall group into which many difficult cases fall. They are distinctive for their context dependency, and for not being like the first group, but beyond that it is hard to find generalizations to tie them together.

Hence, I consider the first group of gestures the "easy group" for recognition. Despite the fact that the system did not do particularly well recognizing the first group, there was and is reason to believe that a system or a person might be able to classify a group 1 gesture before understanding it. On the other hand the second group is a "hard group." A human must

understand a gesture of the second group before classifying it – the problem is AI complete. An automatic system (absent domain or context specific information) can best classify a group 2 gesture by what it is not.

However, builders of interactive systems need not entirely despair, because the constraints of domain and context can be powerful enough to limit the group 2 gestures into a small and more easily understood subset. For example, Bolt's Put That There system [8] used natural domain constraints to limit group 1 gestures to deictics, and group 2 gestures to a few simple classes such as "make it that big," and "rotate it by that much." Similarly, map tasks may constrain the user to making deictics and path gestures. The right choice of domain can reduce the problem of group 2 gestures to tractability.

7.5 Future Work

7.5.1 Phatics - Detecting and acknowledging backchannel feedback

Most feedback in American culture is either paraverbal, such as "mmhmm, uh huh" or head nods. In future work, the vision system might detect changes in orientation associated with a nod, and may not need all the machinery of HMMs to classify them. Human speakers detect backchannel from listeners and use it to decide whether to dwell longer on a topic and explain more fully, or to hurry on to the next topic. REA, when she is developed enough, could do likewise.

7.5.2 Responding to user's conversational style

In future work, the REA system could be programmed to estimate the frequency and size of the user's gestures, and use this information to modulate her own gestural style. It remains to be seen if this will appear to the user as a pleasant adaptation or an annoying imitation.

7.5.3 Principle of symmetry (agent's input and output)

In future work, we plan to make REA's inputs and outputs symmetric. There will be both an external symmetry, such that REA will generate and express the same behaviors that she inputs and processes. More importantly, there will be an internal symmetry, so that the same representations and semantics will be used to code and process inputs and outputs. Thus the external symmetry will occur as a natural consequence of the internal semantic symmetry.

Bibliography

[1] Minoru Asada and Saburo Tsuji. Representation of three dimensional motion in dynamic scenes. *Computer Vision, Graphics, and Image Processing*, 21(1):118–144, Jan 1983.

[2] 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.

[3] Ali Azarbayejani, Christopher Wren, and Alex Pentland. Real-time 3-D tracking of the human body. In *Proceedings of ImageCom96, 3rd International Conference Communicating by Image and Multimedia*, pages 19–24, Bordeaux France, May 1996. Union Europeenne de Radio-Television / European Broadcasting Union.

[4] C. D. Barclay, J. E. Cutting, and L. T. Kozlowski. Temporal and spatial factors in gait perception that influence gender recognition. *Perception and Psychophysics*, 23(2):145–152, 1978.

[5] Janet B. Bavelas, Nicole Chovil, Douglas A. Lawrie, and Allan Wade. Interactive gestures. *Discourse Processes*, 15:469–489, 1992.

[6] A.F. Bobick, S.S. Intille, J.W. Davis, F. Baird, L.W. Campbell, Y. Ivanov, C.S. Pinhanez, A. Schütte, and A. Wilson. The KidsRoom: A perceptually-based interactive and immersive story environment. *"PRESENCE: Teleoperators and Virtual Environments"*, 8(4):367–391, August 1999.

[7] Bobick, A. and J. Davis. Real time recognition of activity using temporal templates. In *IEEE Workshop on Applications of Computer Vision*, Sarasota, December 1996.

[8] Richard A. Bolt. *The Human Interface: Where People and Computers Meet.* Lifetime Learning Publications, Belmont, CA, 1984.

[9] 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. MIT Media Laboratory, IEEE Computer Society Press.

[10] Justine Cassell. *A Framework for Gesture Generation and Interpretation.* in *Computer Vision in Human-machine Interaction*, R. Cipolla and A Pentland eds, Cambridge University Press, Cambridge, U.K., 1998.

[11] Justine Cassell, Timothy W. Bickmore, Mark Billinghurst, Lee W. Campbell, Kenneth Chang, Hannes Hogni Vilhjalmsson, and Hao Yan. Embodiment in conversational interfaces: REA. In *ACM CHI99 Conference on Human-Computer Interaction*, pages 520–527, 1999.

[12] Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. Animated conversation: Rule-based generation of facial expression, gesture, and spoken intonation for multiple conversational agents. In *Computer Graphics (ACM SIGGRAPH 1994 Proceedings)*, pages 413–420, July 1994.

[13] Claudette Cedras and Mubarak Shah. A survey of motion analysis from moving light displays. In *Proc. 1994 IEEE Conf. on Computer Vision and Pattern Rec.*, pages 214–221. IEEE Press, 1994.

[14] C.J. Cohen, L. Conway, and D. Koditschek. Dynamical system representation, generation, and recognition of basic oscillatory motion gestures. In *Second International Conference on Face and Gesture Recognition*, pages 60–65, Killington VT, Oct 2000. IEEE Computer Society Press.

[15] Y. Cui and J. Weng. Learning-based hand sign recognition. In *Proc. of the Intl. Workshop on Automatic Face- and Gesture-Recognition*, pages 201–206, Zurich, 1995.

[16] Y. Cui and J.J. Weng. Hand sign recognition from intensity image sequences with complex backgrounds. In *Second International Conference on Face and Gesture Recognition*, pages 88–93, Killington VT, Oct 1996. IEEE Computer Society Press.

[17] James W. Davis and Aaron F. Bobick. The representation and recognition of action using temporal templates. In *Proceedings Computer Vision and Pattern Recognition (CVPR'97)*, pages 928–934, June 1997.

[18] Niels Diffrient, Alvin R. Tilley, and Joan C. Bardagjy. *Humanscale*, chapter 1. MIT Press, Cambridge, MA, 1974.

[19] Winand H. Dittrich. Action categories and the perception of biological motion. *Perception.*, 22(1):15, 1993.

[20] David Efron. *Gesture and Environment*. Kings Crown Press, New York, NY, 1941.

[21] Paul Ekman and Wallace Friesen. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1(1):49–98, 1969.

[22] Tim Finin, Richard Fritzson, Don McKay, and Robin McEntire. The KQML information and knowledge exchange protocol. In *Third International Conference on Information and Knowledge Management (CIKM-94)*, pages 456–463, Gaithersburg, MD, 1994. ACM Press. see `http://umbc.edu/ finin/papers/` and `http://www.cs.umbc.edu/kqml/`.

[23] D. M. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding: CVIU*, 73(1):82–98, 1999.

[24] S. Gutta, J. Huang, I.F. Imam, and H. Wechsler. Face and hand gesture recognition using hybrid classifiers. In *Second International Conference on Face and Gesture Recognition*, pages 164–169, Killington VT, Oct 1996. IEEE Computer Society Press.

[25] D. D. Hoffman and B. E. Flinchbaugh. The interpretation of biological motion. *Biological Cybernetics*, 42:195–204, 1982.

[26] T. Huang and V. Pavlovic. Hand gesture modeling, analysis, and synthesis. In *International Workshop on Automatic Face and Gesture Recognition*, pages 73–79, Zurich, 1995.

[27] Gunnar Johansson. Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, 14(2):201–211, 1973.

[28] Gunnar Johansson. Spatio-temporal differentiation and integration in visual motion perception. *Psychol. Res.*, 38:379–393, 1976.

[29] Adam Kendon. *Some Relationships Between Body Motion and Speech: An Analysis of an example*, pages 177–210. Pergamon Press, Elmsford, NY, 1972.

[30] Adam Kendon. *Gesticulation and Speech: Two Aspects of the Process of Utterance*, pages 207–227. Mouton, The Hague, 1980.

[31] David B. Koons. Capturing and interpreting multi-modal descriptions with multiple representations. In *Proc. of the AAAI Symposium on Intelligent Multi-Modal Systems*. AAAI Press, March 1994.

[32] Robert M. Krauss, Palmer Morrel-Samuels, and Christina Colasante. Do conversational hand gestures communicate? *Journal of Personality and Social Psychology*, 61:743–754, 1991.

[33] R. Liang and M. Ouhyoung. A real-time continuous gesture interface for taiwanese sign language. In *Submitted to UIST*, 1997.

[34] D. McNeill, J. Cassell, and E. Levy. Abstract deixis. *Semiotica*, 95-1/2:5–19, 1993.

[35] David McNeill. *Hand and Mind*. University of Chicago Press, Chicago, IL, 1992.

[36] Erin Marie Panttaja. Recognizing intonational patterns in english speech. Master's thesis, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge MA, May 1998.

[37] V. I. Pavlovic, J. M. Rehg, and J. MacCormick. Impact of dynamic model learning on classification of human motion. In *Proc. 1994 IEEE Conf. on Computer Vision and Pattern Rec.*, Hilton Head, SC, June 2000. IEEE.

[38] V. I. Pavlovic, R. Sharma, and T. S. Huang. Visual interpretation of hand gestures for human-computer interaction: a review. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):677–95, July 1997.

[39] Vladimir Pavlovic, Brendan Frey, and Thomas Huang. Time series classification using mixed-state dynamic bayesian networks. In *CVPR'99*, Ft. Collins, CO, 1999.

[40] F. Quek, D. McNeill, R. Ansari, X. Ma, R. Bryll, S. Duncan, K-E. McCullough, and C. Kirbas. Gesture cues for conversational interaction in monocular video. In *International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems (ICCV RATFG-RTS'99)*, pages 119–126, Corfu Greece, Sept 1999.

[41] F. Quek, D. McNeill, R. Bryll, C. Kirbas, H. Arslan, K-E. McCullough, N. Furuyama, and R. Ansari. Gesture, speech, and gaze cues for discourse segmentation. In *Computer Vision and Pattern Recognition (CVPR 2000)*, Hilton Head SC, June 2000.

[42] F.K.H. Quek and M. Zhao. Inductive learning in hand pose recognition. In *Second International Conference on Face and Gesture Recognition*, pages 78–83, Killington VT, Oct 1996. IEEE Computer Society Press.

[43] 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.

[44] R. Rashid. Towards a system for the interpretation of moving light displays. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2:574–581, 1980.

[45] Rehg, J. and T. Kanade. Model-based tracking of self-occluding articulated objects. In *Proc. Fifth International Conference on Computer Vision*, pages 612–617, Boston, MA, 1995.

[46] W. T. Rogers. The contribution of kinesic illustrators towards the comprehension of verbal behavior within utterances. *Human Communication Research*, 5:54–65, 1978.

[47] Ross Cutler and Matthew Turk. View-based interpretation of real-time optical flow for gesture recognition. In *Third International Conference on Automatic Face and Gesture Recognition*, pages 416–421, Nara, Japan, April 1998.

[48] J. Schlenzig, E. Hunter, and R. Jain. Recursive identification of gesture inputs using hidden markov models. *Proc. 2nd Ann. Conf. on Applications of Computer Vision*, pages 187–194, December 1994.

[49] Jeffrey Mark Siskind. Visual event classification via force dynamics. In *Proceedings AAAI-2000*, pages 149–155, August 2000.

[50] Carleton J. Sparrell. Coverbal iconic gesture in human-computer interaction. Master's thesis, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge MA, June 1993.

[51] T. Starner, J. Weaver, and A. Pentland. A wearable computer based american sign language recognizer. In *First Intl. Symp. on Wearable Computing*, pages 130–137, Cambridge, MA, 1997. IEEE Press.

[52] T. E. Starner. Visual recognition of american sign language using hidden markov models. Master's thesis, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge MA, February 1995.

[53] 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.

[54] Kristinn Thorisson. *Communicative Humanoids*. PhD thesis, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge MA, July 1996.

[55] Kristinn Thorisson, David B. Koons, and Richard A. Bolt. Multi-modal natural dialogue. In *Proceedings of CHI'92 ACM conference on Human Factors*, pages 653–4. ACM Press, May 1992.

[56] J. Triesch and C. von der Malsburg. Robust classification of hand postures against complex backgrounds. In *Second International Conference on Face and Gesture Recognition*, pages 170–175, Killington VT, Oct 1996. IEEE Computer Society Press.

[57] J.A. Webb and J.K. Aggarwal. Structure from motion of rigid and jointed objects. *Artificial Intelligence*, 19:107–130,, 1982.

[58] Alan D. Wexelblat. A Feature-Based Approach to Continuous-Gesture Recognition. Master's thesis, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge MA, May 1994.

[59] A. Wilson and A. Bobick. Recognition and interpretation of parametric gesture. In *International Conference on Computer Vision (ICCV'98)*, pages 329–336, 1998.

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

[61] A. D. Wilson, A. F. Bobick, and J. Cassell. Recovering the temporal structure of native gesture. In *Second International Conference on Face and Gesture Recognition*, pages 66–71, Killington VT, Oct 1996. MIT Media Laboratory, IEEE Computer Society Press.

[62] Christopher R. Wren, Brian P. Clarkson, and Alex P. Pentland. Understanding purposeful human motion. In *Fourth International Conference on Face and Gesture Recognition*, Grenoble FR, Oct 1996. MIT Media Laboratory, IEEE Computer Society Press.

[63] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. In *Proc. 1992 IEEE Conf. on Computer Vision and Pattern Rec.*, pages 379–385. IEEE Press, 1992.