Coverbal Iconic Gesture in Human-Computer Interaction

by

Carlton James Sparrell

S.B. Electrical Science and Engineering
Massachusetts Institute of Technology
Cambridge, Mass.
1988

SUBMITTED TO THE PROGRAM IN MEDIA ARTS AND SCIENCES,
SCHOOL OF ARCHITECTURE AND PLANNING
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 1993

Signature of Author_____
Program in Media Arts and Sciences
May 7, 1993

Certified by_____
Richard A. Bolt
Senior Research Scientist, MIT Media Laboratory
Thesis Supervisor

Accepted by_____
Chairman, Departmental Committee on Graduate Students

Coverbal Iconic Gesture in Human-Computer Interaction

by

Carlton James Sparrell

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning
on May 7, 1993, in partial fulfillment of the
requirements for the degree of
Master of Science

Abstract

This thesis discusses the recognition and application of *Coverbal Iconic Gesture* in *Human-Computer Interaction.* Coverbal Iconic Gesture is a class of gesture where the hand's configuration and movement becomes an iconic representation of some concrete object referred to simultaneously in speech. A method of gesture recognition is proposed in which the raw data from a hand tracking device is abstracted into an intermediate descriptive representation that can be interpreted simultaneously with the natural language input from a speech recognition system. A prototype *multi-modal natural language* system is described, which has been implemented as a platform for demonstrating the methods detailed in this work.

Thesis Supervisor: Richard A. Bolt
Title: Senior Research Scientist, MIT Media Laboratory

The author acknowledges the support of the Defense Advanced Research Projects Agency under Rome Laboratories Contracts F30602-89-C-0022 and F30602-92-C-0141.

The views and conclusions contained in this document are those of the author and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the U.S. Government.

Thesis Reading Committee

Read by____

Aaron F. Bobick
Assistant Professor of Computational Vision
MIT Media Laboratory

Read by____

Glorianna Davenport
Assistant Professor of Media Technology
MIT Media Laboratory

Read by____

Brenda K. Laurel
Research Artist
Interval Research Corporation

Contents

List of Figures

Acknowledgments

To my parents, Betty and Jim, for their patience and support through both my undergraduate and graduate careers at this institution where they first met.

I would like to thank my advisor, Dr. Richard Bolt, for his inspiration and vision of *the human interface*. I am also grateful to my readers, Glorianna Davenport, Brenda Laurel, and Aaron Bobick.

Dave Koons has my extreme gratitude for sharing his thoughts, ideas, and his time. He also deserves credit for the speech interpreter and representational spaces, which allowed my gestures to be coverbal.

I would also like to acknowledge the other AHIG grad students, Kristinn Thorisson, Alan Wexelblat and Edward Herranz. Kris gave me extensive editorial help, Alan helped out proofreading and Ed spent many hours helping me get familiar with all the AHIG toys.

Several undergraduate researchers have contributed to this work. David Berger developed the dispatcher which facilitated the distributed network processing. I thank him for the long hours he spent helping me out, even after he had finished his tenure here. Chris Wren provided me with *MVERSE*, and was always available to help out as a UNIX guru. Thanks go to Greg Spurrier for implementing the *Object Knowledge Base*, and finding 101 things to do with an animated dog. Brian Brown put in many hours as my right-glove man, producing a gesture recorder and experimenting with many of my early theories.

Special thanks to Susan Dunbar, for the photographs that led to Figures 2-2 and 4-4. I am also grateful for her encouragement of my creative side, which the Media Lab has given me the opportunity to explore.

Finally, thanks to my many close friends in and around the Boston area and beyond. It has been with their continuing support that I have been able to work hard and play hard, while still maintaining my sanity.

Notices

The following trademarks are referred to throughout this document:

- DataGlove is a registered trademark of VPL Research, Inc.
- HARK is a registered trademark of BBN Systems and Technologies, a division of Bolt Beranek and Newman, Inc.
- Starbase and Real Time Interface are registered trademarks of Hewlett Packard.
- 3Space is a registered trademark of Polhemus Navigation Sciences.

Chapter 1

Introduction

1.1 Inspiration

> Two students sit in front of a computer screen. One student is describing an animation. With a few short phrases and some careful hand gestures, she describes the layout of a synthetic room, the motion of a synthetic actor, and the camera angle. The other student watches her gestures, listens to the scene description, and then types away furiously at the keyboard. After entering a series of graphics commands, the animation starts to unfold on the computer screen. The first student watches carefully, and then starts to describe changes that might improve the effort. [1]

People often take advantage of gesture when describing objects, locations and actions. The scenario listed above shows such an example. When one student describes a scene to the other, it is with the use of coverbal gesture: gesture accompanied by speech. The second student has to translate this mode of communication into a sequence of graphics commands and approximate coordinate values. In this sense, the student at the keyboard

[1] The scenario described is based on an actual experience using the CINEMA system for the MIT graduate course *Cinematic Knowledge and Synthetic Cinema* taught spring semester 1992 by Glorianna Davenport and David Zeltzer. (see Drucker, et al. [1992]).

is acting solely as an interface agent. Ideally, the first student should be able to describe the scene to the computer directly, complete with both speech and gesture.

This work presents an approach for designing a computer interface that can interpret some forms of coverbal gesture; in particular, those gestures where the hand becomes an iconic representation of a concrete object described in accompanying speech. Gesture recognition in this sense represents a departure from previous efforts, which were concerned mainly with using the hand as a direct manipulation device or as a means for entering symbolic commands.

1.2 Coverbal Iconic Gesture

When talking face-to-face with one another, people commonly gesture along with their words. Often these gestures provide images that cannot be easily described using speech. There is evidence that gesture and speech are combined in an integrated system of communication. Graham and Heywood [1976] showed that the elimination of hand gestures by the speaker had a detrimental effect on speaking performance. Similarly, Graham and Argyle [1975] showed degraded comprehension in listeners who were not able to see a speaker's hands during communication.

The label 'gesture' is often used in a very broad sense. The scope of this work covers only a narrow, but useful class of gesture which I refer to as *Coverbal Iconic Gesture*. 'Coverbal,' though technically redundant with Iconic, refers to gestures accompanied by speech. 'Iconic' comes from the McNeill [1992] classification of gesture. McNeill's classes, which I will use throughout this work, are provided below.

Iconic: A gesture that, in it's execution and manner of performance, refers to a concrete event, object, or action that is also referred to in speech at the same time.

Metaphoric: Metaphoric gestures are similar to Iconics in that they present imagery, but present an image of an abstract concept, such as knowledge, language, the genre

of the narrative, etc.

Deictic: Deictic gestures are pointing movements, which are prototypically performed with the pointing finger, although any extensible object or body part can be used including the head, nose, or chin, as well as manipulated artifacts.

Beats: Movements that do not present a discernible meaning. Beats are typically small, low energy, rapid flicks of the fingers or hand.

1.3 Iconic vs. Symbolic Gesture

McNeill's classification falls solely under the 'gesticulation' end of what he refers to as Kendon's Continuum. Specifically this gesticulation is defined as "idiosyncratic spontaneous movements of the hands and arms accompanying speech."

The Kendon's Continuum (McNeill [1992]):

Gesticulation ⇒ Language-like Gesture ⇒ Pantomimes ⇒ Emblems ⇒ Sign Language.

This continuum shows the broad range of free hand gesture. McNeill points out that while moving from left to right on this scale, "(1) the obligatory presence of speech declines, (2) the presence of language-like properties increases, and (3) idiosyncratic gestures are replaced by socially regulated signs."

Most current methods of gesture recognition have targeted the Sign Language end of the continuum. This is true whether the sign language being detected was a standard language such as ASL (Kramer & Leifer [1989]), or a sign language specific to the application (Rubine [1991], Fisher [1986]).

Gesticulation interfaces to date have primarily focused on deictic gestures relying on a mouse (Neil & Shapiro [1991]), a screen pointer (Bolt [1980]), or rigid interpretation of a pointing gesture (Thorisson, Koons & Bolt [1991]). It is no coincidence that these early gesticulation interfaces rely on deictic gesture. Deictic gestures are perhaps the most

common and most important gesture type. It is also the simplest class of gesture to detect, if not necessarily to interpret.

Iconic gesture is the natural next step for gesticulation interfaces. First, it opens up interaction far beyond the pointing of deictic gestures. Second, it is feasible to detect many iconic gestures because they deal with concrete objects that are specifically referred to in the speech channel. Metaphorics and beats, on the other hand, usually have a meaning that is less well defined. [2]

1.4 Why is Coverbal Gesture Recognition Different?

The method of operation of most gesture recognition techniques is to map a hand movement to some rigid interpretation of symbolic meaning. Some of the gestures used in these systems mimick real world actions. A user might move an object by "picking it up" with a "virtual hand" (Fisher, et al. [1986]). Other actions are more arbitrary. A user might use a two handed finger posture to cause a menu to appear, and then use a wrist turn to select an item. Gesture recognition has been done by various methods including statistical pattern recognition for direct manipulation (Rubine [1991]) and Recurrent Neural Networks for sign language recognition (Murakami & Taguchi [1991]). In sign language interfaces to date, gestures recognized are limited to sign alphabets.

Simple direct manipulation interfaces can be easy to learn and to operate. Unfortunately, as the desired repertoire of actions becomes more extensive, the functionality of the system quickly becomes hidden under hierarchies of different modes and menu trees. Another limitation is that the gestures cannot be done "any old way" but in a manner dictated by specific constraints and conventions.

Sign Language interfaces are useful for allowing the deaf to communicate with the sign

[2] Beats could be used as a clue in turn taking. These gestures are often present when a breakdown of the speech channel occurs, as when the speaker is trying to think of the right word to say. Presence of these gestures could be used as a cue that the current stream of speech has not yet concluded.

illiterate. For the novice user, learning to use such an interface would require the learning of 26 (ASL) to 42 (Japanese SL) alphabet signs. These systems also suffer from forcing the user to spell out words rather than use the true lexicons of sign languages for the deaf.

Coverbal gesture recognition should not specifically map symbolic interpretations to a library of hand movements. The meaning of a hand movement or posture is very much dependent on the meaning of the accompanying speech. The speech sets up a context for the gesture and similarly the gesture can alter the meaning of the utterance. Interpretation of this type of interaction requires an integrated approach where the parser can simultaneously look at the clues from the user's speech and gesture to develop a model of the user's intent.

To allow this type of interpretation, I am proposing a representation for the hand actions that is concise, flexible and should preserve much of the naturalness of common iconic gesture. This form is a level of abstraction loosely analogous to word phrases in the speech channel. Building this intermediate representation delays final evaluation until integrated interpretation of the speech and gesture can take place.

1.5 Related Work

Much of this work is based on the research of cognitive scientists who have studied natural human gesture. In particular, the author relied on the research of McNeill [1992], McNeill & Levy [1982], Poizner, Klima & Bellugi [1987], Klima & Bellugi [1979], and Stokoe [1960]. McNeill's work highlights many valuable observations of naturally occurring coverbal gesture. Poizner, Klima and Bellugi concentrate on the discussion of sign language and how it relates to human cognition. Stokoe provides an in depth discussion of sign language primitives. Despite the differences in interpretation of sign language and coverbal gesture, sign language provides some valuable insight into our capabilities to perceive hand postures and movements.

Many gesture recognition systems have been developed since the appearance of whole

hand input devices. Fisher, et al. [1986] demonstrated one of the first such systems as a means of navigating and manipulating a virtual environment. More recently, Kramer & Leifer [1989] demonstrated the recognition of ASL finger spelling. Rubine [1991] developed a control system using learned 'free-hand drawing' gestures (i.e. gestures specified by the path drawn in a 2D plane by the hand or fingers). Sturman [1992] explored whole hand input in general and his thesis provides a valuable survey of gestural interfaces.

Butterworth [1992] extended Fisher's idea of using hand input to interact with a virtual world, implementing a virtual environment system called 3DM. This system, likened by the author to a 3D MacDraw, used a 3D mouse instead of hand input, to control a palette of command items.

Bolt [1980] demonstrated the use of combined speech and pointing in a system called "Put-That-There." This example used a 3-D space sensing cube to generate deictic references. More recently, Thorisson, Koons & Bolt [1991] demonstrated combined speech, deictic hand gestures, and gaze. Neal and Shapiro [1991] demonstrated a similar system, only with mouse and keyboard input. Herranz [1992] demonstrated the usefulness of two hand input, together with eye tracking. His system allowed a user to scale, rotate and move objects by using speech and two handed symbolic gesture.

1.6 Prototype Overview

The demonstration platform for the ideas in this thesis is the Virtual Environment Coverbal Iconic Gesture interface system (VECIG). This interface allows a user to interact with a three-dimensional computer generated environment through the use of speech and accompanying gestures. The specific interaction scenario depends on the object descriptions loaded into the Object Knowledge Base and the lexicon present in the interpreter parse code. Currently, a scenario has been designed to allow a user to arrange a room with furniture and other simple objects. Iconic gesture types supported allow relational object placement, object manipulation, and simple object animation. The specific details of this

system are described in chapter 4.

Chapter 2

Foundations

The goal of this work is to develop a coverbal gesture recognition scheme that will allow interaction to be as natural as possible. By natural, I refer to the use of commonly occurring, spontaneous speech and gesture. The naturalness of any system of this sort is limited by its robustness, and producing a catch-all scheme is well beyond the scope of this work. It is a goal, however, to establish a foundation on which more robust systems can be built. To accomplish this goal, it is important to look at how naturally occurring iconic gestures might be described and interpreted.

The previous chapter defined Iconic Gesture and specified how this class fits into the continuum of gestural communication. Iconic gesture can be further broken down into three subgroups. The names of these subgroups vary by author. I will use as a standard the groups proposed by Rimé and Schiaratura [1991], *spatiographic*, *kinetographic*, and *pictographic*. The relation of these sub-groups, combined with the McNeill hierarchy of gestures is shown in Figure 2-1.

Spatiographic iconic gestures are used to represent some spatial relationship of the referents. This is typically done by using the relative location of gestures in the speaker's reference space. An example of this might be someone explaining the layout of a parcel

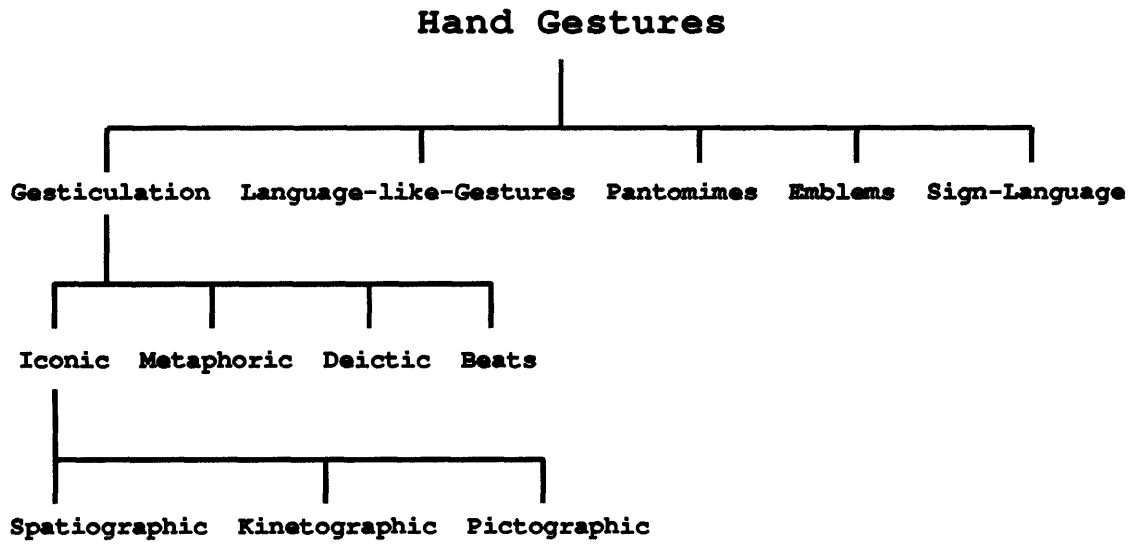

Figure 2-1: The Gesture Hierarchy

of land, "The house was over [gesture] here, and the barn was over [gesture] here." [1]

Pictographic iconic gestures are used to present the shape of an object. An example of this might be someone showing the outline of an object with their two hands, "The vase was shaped like [gesture] this."

Kinetographic iconic gestures are used to represent some type of motion. An example of this might be someone explaining an accident that they witnessed, "The car swerved off like [gesture] this..."

There are three parameters among these gesture types that contain the meaning. The first parameter is spatial information. When describing the relative location of objects, the speaker sets up a reference space in front of themself. Once a point of reference has been established in this space, subsequent gestures provide a spatial relationship in this gesture space.

A second parameter is that of hand configuration (shape) and orientation. With iconic

[1] As will be discussed later, it has been observed that gestures precede speech in meaning, with the stroke phase terminating just before or at the peak syllable.

gesture the hand shape often provides an image of that object. The hand may become the object, such as a flat hand representing a car, or a curled hand representing a spider. The hand may also mime the grasping of the object, such as a curved hand holding a glass. The orientation provides additional information, answering such questions as how the barn is oriented with respect to the house, or illustrating the direction a car is headed.

The third parameter is that of hand motion. The path that the hand draws out might represent the movement of the car, or portray a falling body. The path might also illustrate the outline of an object. Hand motion is also crucial in determining the phase of the gesture, as will be discussed later.

2.1 The Search for Primitives

Each of the hand parameters can assume a wide array of values, yet the parametric information needed to interpret a gesture can typically be represented by very few values. In a spatiographic gesture, for example, we may only need to know the place of the hand at one moment in time to disambiguate the speech. The task of interpreting the gesture then becomes a problem of analyzing the speaker's hand movements, in the context of the speech, in an effort to zero in on the important information. The significance of this is that the bulk of the coverbal-gesture recognition process can be performed on qualitative descriptions of the hand movements, and only a small amount of parametric information need be retained. This section proposes a set of device-independent gesture primitives that may be used to describe gestures for the purpose of coverbal interpretation.

Ideally, a set of primitives is one that is mutually exclusive and collectively exhaustive. Mutually exclusive in this case can be thought of as gesture features that are sufficiently differentiable to alter the perceived meaning. Developing a collectively exhaustive set is a bit more of a task. Iconic gesture is by definition spontaneous and idiosyncratic, leaving potentially endless possible variations. In reality, however, we find that a small set of particular features can account for the majority of iconic gestures.

Despite the vast differences between Gesticulation and Sign Language, McNeill observed that hand shapes in spontaneous gesture can be approximated to those in ASL. One explanation for this is that sign languages typically evolve out of spontaneous gesture. The significance of this observation is that it suggests that primitives for spontaneous gesture might be found by examining the research of those whom have studied ASL closely. The ASL classification was first developed by Stokoe [1960] and his work is summarized in Klima and Bellugi [1979]: "Stokoe observed that ASL signs are not just uniquely and wholly different from one another and posited that they can be described in terms of a limited set of formational elements." Stokoe's Dictionary of American Sign Language lists 19 hand configurations, 12 placement primes and 24 primes of movement. These primes, which he calls cheremes, in analogy to phonemes, are further broken down into various sub-primes called allochers.

2.1.1 Configuration Primes

McNeill and Levy [1982] performed an experiment in which subjects were shown an animated cartoon and asked to retell the story to another person. During the recounting, the speaker's gestures were recorded and then cataloged. The researchers were able to categorize every hand shape used among all six narrators using 20 of Stokoe's primitive hand configuration sub-primes. McNeill warns that narratives of other genres would undoubtedly produce a different distribution of primitives. Accordingly, non-narrative input would yield different results as well. This study does suggest, however, that configuration primes for natural gesture might be taken from the list of ASL primes, and that a small number of hand configuration primitives is sufficient, given a specific context. The specific list of primes used for the current work is detailed in chapter 4.

Orientation of the hand can be considered a subclass of configuration. Although Stokoe does not present orientation primes, a general set of orientation primes would provide valuable information when analyzing hand input data. The primes used in this work distinguish the orientation of the palm by quantizing the tangent vector out of the front

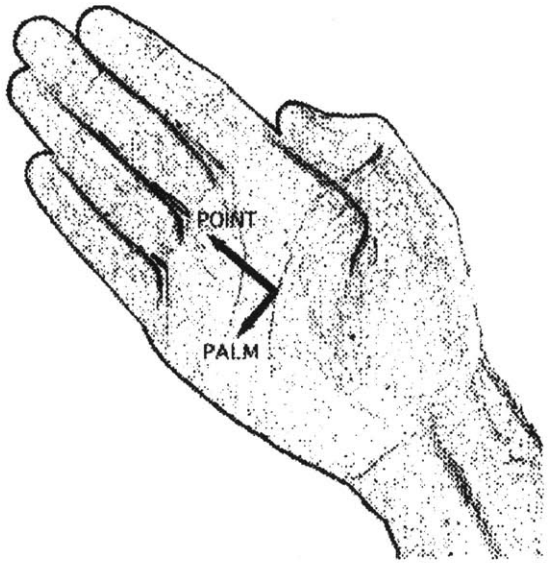

Figure 2-2: Hand Orientation Vectors

of the hand and normal vector out of the palm as forward, back, left, right, up, or down. Henceforth, the tangent will be referred to as the *point* vector, and the normal will be referred to as the *palm* vector. The precise vectors are illustrated in Figure 2-2.

2.1.2 Place of Articulation Primes

While ASL research provides an adequate foundation of configuration primes, the list of place markers is less appropriate. Stokoe's place of articulation cheremes primarily reference parts of the body. He specifies six place markers around the face and head, four on the arms, one for the trunk and one away from the body for a pause or rest.

Spontaneous gesture, by contrast, almost exclusively occurs in the space out in front of the speaker's body. McNeill observed that gestures in adults are usually performed in a space roughly defined by a shallow disk in front the speaker. His study recorded gesture placement and mapped these places of articulation into the subdivisions shown in Figure 2-3. Almost all iconic gestures recorded fell inside the periphery, center, or

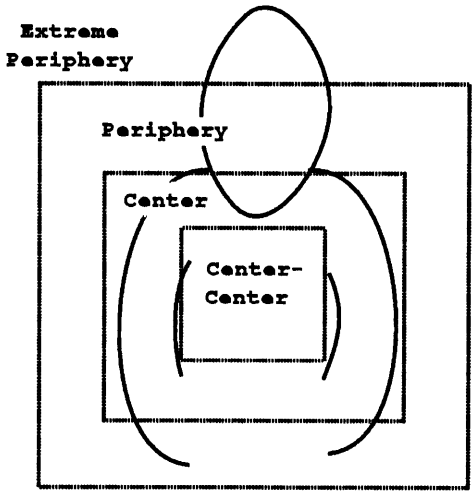

Figure 2-3: McNeill's Subdivision of Gesture Space

center-center regions. The distribution within these regions was almost uniform. Ideal place of articulation primes should only divide this space into regions that would affect the gesture meaning. I will distinguish between the extreme periphery, most notably the "rest space" below the center, and the more central regions which I will henceforth refer to as the gesture space. For iconic gesture, I suggest the set of place markers consist solely of a binary parameter specifying whether the hand was in gesture space or not.

More robust gesture schemes would benefit from a more exact set of place of articulation primes. McNeill's studies suggest that metaphorics tend to be articulated more specifically in the lower center region. Deictics more often show up in the periphery, and beats often are bunched in a small off center region that is idiosynchratic to the speaker. These distinctions can be of importance in trying to filter out certain gesture types, or in trying to classify the gesture type.

2.1.3 Movement Primes

Spontaneous gesture movements suffer from being highly context specific, but common movement types similar to those seen in ASL do emerge. Stokoe [1960] specifies the

distinct linear hand movements of up, down, left, right, to, and fro. These movements are also common in various phases of spontaneous gesture and the distinct meanings make them excellent primes.

Stokoe lists the reflective movements of forward then back, left then right, to then fro, etc. as movement cheremes. While these movements do occur within normal gesticulation, they are components of different phases of a gesture and will be treated as combinatoric constructions of the other primes. The wrist cheremes in ASL also map well into coverbal gesture. These consist of the supination (wrist bending forward), pronation (wrist bending back), and rotation clock- and counter-clockwise.

Circular, spiral, and other complex motions of the hands are observed in everyday gesture. Of these, only circular exists as a Stokoe prime. Identifying these more intricate motions would take significant observation of test subjects and is outside the scope of this work.

2.2 Gesture Parametrics

The gesture primitives can be used to transform the raw data from hand tracking devices into a description that can be parsed in the context of the accompanying speech. The primitives can also be used to greatly reduce the amount of data that requires processing. Sequential data records may be compared, and only those that represent a change from the previous record need be extracted for interpretation. Parametric vectors representing the hand position and orientation can be extracted along with these distinct feature records. These vectors give more detailed information of the hand at significant points in time; those points where the descriptive primitives changed.

The final step of interpretation of a spatiographic gesture is to map a specific location and sometimes orientation to the object represented by the hand. When a feature record has been determined to be the point of reference, the location and orientation vectors at that moment in time will contain the only necessary parametric information in the entire gesture. With kinetographic and pictographic gestures, the parametric vectors will provide

a keyframe representation of the hand movement for the record or records determined significant by the feature descriptions. The process for determining the significant records is discussed in Chapter 3.

2.3 The Timing of Gesture

Timing of coverbal gesture is of utmost importance in understanding its meaning. There are two aspects of this timing that must be relied on for iconic gesture recognition: gesture phase and gesture synchrony. The phase of the gesture can be determined by looking at the the movement with respect to the preceding and subsequent gesture segments, as discussed below. The synchrony of the gesture with the speech will determine what part of the speech gives the context.

2.3.1 Phases of Gesture

Kendon [1980] specified that gestures consist of at least one, usually more phases of movement. These phases were classified as *preparation, pre-stroke hold, stroke, post-stroke hold,* and *retraction.* Of these phases, only the stroke is obligatory. The preparation phase almost always occurs to some extent, though the retraction may be dropped if another gesture follows immediately. Most gestures occur by themselves with the hand returning to rest before another is formed. The hold phases occur whenever necessary to regain synchronization of speech and gesture (McNeill [1992]).

The stroke of the gesture is where the information lies. In iconic gesture, this phase will usually occur near the center of gesture space and is synchronized with the accompanying speech. The other phases of the gesture are useful bookends to help isolate the stroke phase.

2.3.2 The Synchrony of Gesture

The success of coverbal gesture recognition depends heavily on the timing relationship between the stroke phase of the gesture and the linguistic content of the utterance. In iconic gesture this means the gesture will be synchronized with discussion of a concrete object or action. Specifically, McNeill [1992] has observed that: "Gestures are integrated with the speech output. They anticipate the speech in their preparation phase, and synchronize with it in the stroke phase, which is timed to end at or before, but not after, the peak syllable."

In multi-modal interpretation, this tightly matched synchrony allows a tie between the linguistic model of the speech and the spatial model portrayed by the hands. When building a prepositional relation, for example, the gesture stream can be examined near a specific time for any spatiographic iconics to provide specific placement information.

Chapter 3

Knowledge Representation and Communication

Chapter 2 discussed the foundations of a feature-based gesture recognition scheme developed from the observations of cognitive scientists studying spontaneous gesture. This feature-based representation gives us a working model of gesture that can be combined with the other modes of a user's input to determine an appropriate response. To interpret this representation, it is necessary to examine what types of information make up our knowledge and look at how we communicate this information through speech and gesture.

3.1 Representation of Information

Hutchins, Hollan & Norman [1986] discuss two terms in relation to human-computer interaction, *The Gulf of Execution*, and *The Gulf of Evaluation.* The gulf of execution refers to the the chasm we must cross to to communicate our mental image of a task we want accomplished to the computer. The gulf of evaluation refers to the return trip of information, where the computer must reflect to us what has been transpired, and allow

us to understand the new state of the system.

In direct manipulation interfaces, this problem is often solved by allowing only small, incremental changes and reflecting them in WYSIWYG [1] fashion. Natural language interfaces, on the other hand, seek to minimize the small incremental details in communication and allow more general interactions where the details are left to the system.

In a virtual environment, for example, a user might desire to place a chair in the corner of a room. With direct manipulation the user would have to create the object and then specify the location and orientation by interacting with some input device. With natural language, the user may only need to state "Place a chair in the far corner of the room." For this to work satisfactorily, there must be a large amount of shared knowledge between the user and the system. The computer must have enough embedded knowledge to act in a manner consistent with the user's expectations. In this example, the computer must know which corner is the far corner in relation to the viewer, and should also know enough to place the chair feet-down on the floor.

3.1.1 Representation of Knowledge for Multi-Modal Interaction

Koons [1993] proposes a representational system for multi-modal descriptions. Elements in this system are described in three interconnected encoding formats, allowing for visual information (such as shape and appearance), metric information (spatial and temporal relations), and algebraic information (categorical or abstract relations). As information is received from the various modes, frames are added to the system containing the new information. Empty slots are then compared against the available information in an attempt to build a complete model of the user's input. When a model is completed, an associated action for that model type is carried out and any new information is added to the visual, metric and algebraic information networks. I will use the term *semantic model* to refer to this generated model of the user's meaning.

[1] "What You See Is What You Get"

This representational system is especially appropriate for multi-modal interactions because each mode is well suited to communicating only one, or sometimes two of the fundamental information types. Speech may carry elements of each of the three fundamental information types, but speech often breaks down in communicating visual and some metric information. Gesture is rarely used to carry algebraic information, but is often well suited to indicate visual and metric information. Other modes, such as gaze, only supply spatial information.

3.1.2 A Speech-Only Example

A simple speech-only example of the construction of the semantic model is illustrated in Figure 3-1. Here the phrase "Make a red cube" is spoken. As each word comes in from the speech channel, frames are generated for each element and combined into component phrases. The word "make" is tied to a verb frame which holds other slots such as subject, location of subject, and reference object. Each of these slots is initially empty. The word "a" is initialized into a determiner frame, "red" is initialized into a adjective frame, and "cube" is initialized into a noun frame.

After the noun frame is created, the complete noun phrase can be generated. All adjectives, for example, are combined into an object description (here just "red"). The noun frame is then combined with the object description and this structure is combined with the determiner. The resulting structure is the noun phrase. This frame now replaces the individual components in the speech buffer. The category slot is set to "red cube" and the number slot is set to one (from the indefinite determiner "a"). Lacking any input from other modes, the spatial slot is still empty.

The noun phrase could be automatically combined with the verb (or verb phrase) here, but this action is delayed until it is determined that no other information will be following. The speaker could be saying "Make a red cube next to the blue cylinder" for example. There are many clues in multi-modal communication that could be used to determine that the phrase is completed, such as a suitable pause, characteristic speech intonation,

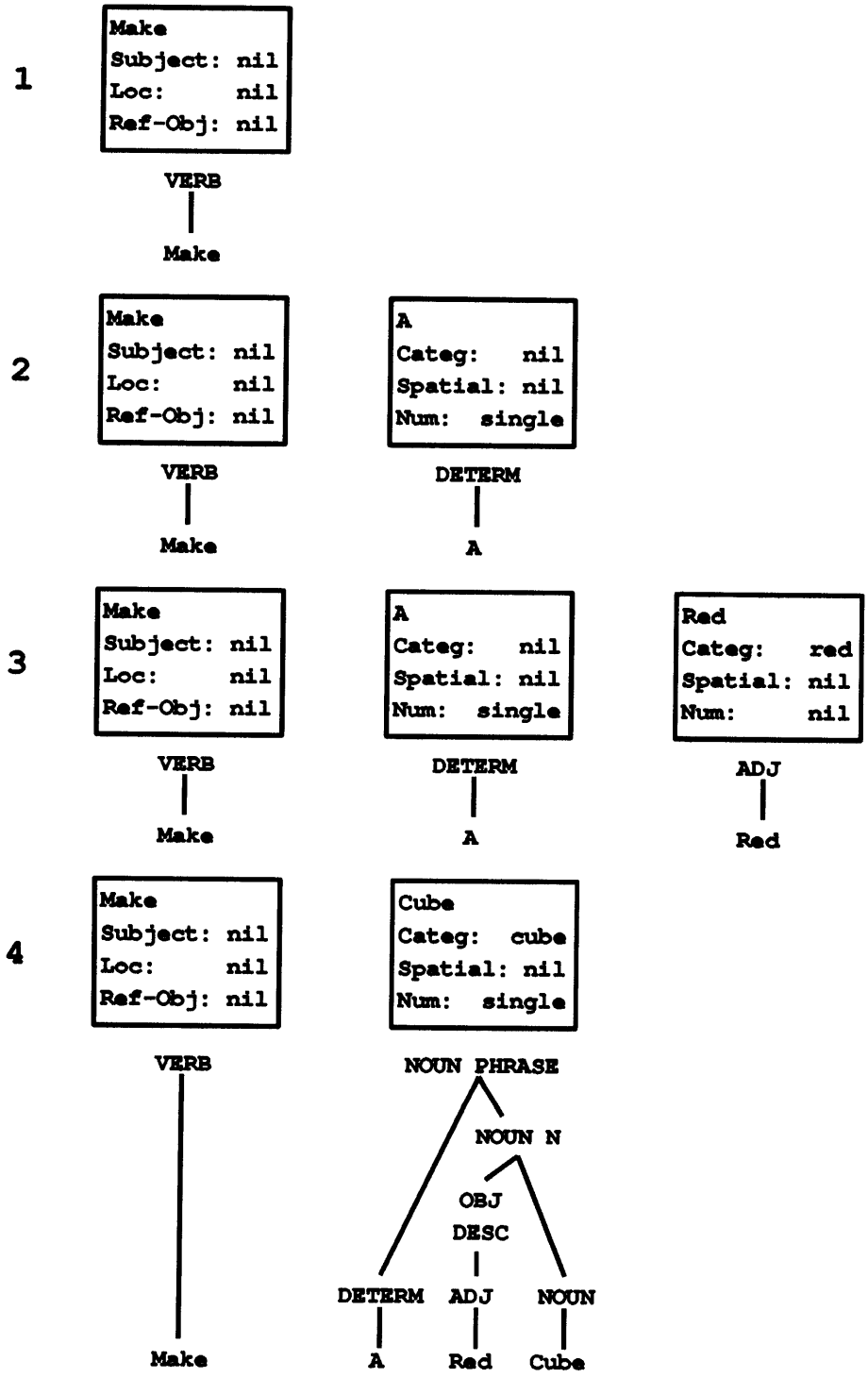

Figure 3-1: Semantic Construction

the speaker's gaze behavior, or absence of gesticulation (Thorisson [1993]). In the demonstration system described in this document, an acceptable pause determines the end of the input phrase.

When the end of input has been established, the verb and noun phrase can be combined and a task method associated with the verb will be carried out. In the above example, a red cube of default size and orientation might be created at some default place in a 3-D environment. The visual, metric, and algebraic information are then entered into the stored information network for future reference.

3.1.3 Bringing in the Hands

Building a model out of speech requires each word be classified so that an appropriate frame can be constructed. Often the interpretation of the word token is dependent on the context in which it is spoken. Although multiple frames and models might co-exist, only one model will be successfully completed.

Creating model elements out of the gesture channel is more of a challenge. There is no predefined grammar for iconic gesture. To determine the meaning of a user's hand movements, one must draw from the context of the speech and rely on knowledge of the objects, actions, or spatial situations being described. The remainder of this chapter proposes a methodology that can be used to extract meaning from gesture in simple interactions.

3.2 Finding Meaning in Gesture

Iconic gestures by definition make reference to concrete objects. Combined with speech, these gestures convey information about where an object is, how an object moves, what an object looks like, or what transformation that object goes through. The computer interface must be able to develop a model of the hand movements and transfer this model

to the appropriate object in the computer's representation of the world. Chapter two described a method for modelling a user's hand movement. To apply this model, we must be able to answer three questions:

- When to look at the user's hand movements.
- Where to look in the user's hand motion.
- How to correlate the hand and object.

After these questions have been answered, it will be possible to extract the correct information to transfer the user's model of the object in gesture space to the computer's model of the object in representational space.

3.2.1 When to Look

Timing of gesture is very important. Kendon [1980] and McNeill [1992] have shown that gestures are highly synchronized to the accompanying speech. Kendon's studies revealed that the stroke phase of gesture slightly precedes, or ends at, the peak phonological syllable of the accompanying speech. Further, he determined that when both speech and gesture are present, both must provide the same meaning at the same time.

There are two immediate applications of these principles. First, when looking for a gesture associated with a given reference, only those movements immediately preceding or coinciding with the speech reference need be examined. Second, when an appropriate motion is found in this time window, it is reasonable to expect that the gesture refers to the same thing as the speech.

3.2.2 Where to Look

The timing information limits us to examining a small window of the gesture information. The size of this window depends on the accuracy and synchrony of the hand and

speech data acquisition channels. Within this window there may be more than one hand movement, while we are often only interested in just part of a single movement. To zero in on the exact point or points of interest, we must know where in the phases of gesture the information lies.

With spatiographic gesture, the hands reference a point and sometimes an orientation in space. This information is clearly static, and can be revealed by a snapshot of the hand in gesture space. Gestures of this type usually occur in two phases, *preparation* and *retraction.* The preparation phase of this gesture is also considered the stroke. When two subsequent spatiographic gestures are performed using the same hand, the stroke of the first gesture may be immediately followed by the stroke of the second with no intervening retraction. The reference of this gesture type occurs at the end of the stroke phase. Examining the hand at this point in time reveals the position and orientation information desired.

Consider the example: "Place a [gesture] cylinder next to the [gesture] cube." A typical set of gestures for this interaction would be one hand moving up into gesture space to represent the cylinder, followed by the second hand moving up to represent the cube. Each hand will come to rest in gesture space at about the same time as the word for the represented object is spoken. The moment when the hand comes to rest becomes the moment of interest in the hand movement. The position and orientation can be extracted from the single hand feature at the end of that stroke phase.

The movement and place of articulation features can be used to find the end of the stroke phase. Motions up into the gesture space from rest space or from a previous reference suggest stroke phases of this sort. Each movement in the reference time window can be examined to find the stroke closest to the reference time. The desired position and orientation can then be extracted from the final feature record of that movement.

Kinetographic and pictographic gestures are dynamic. The information for these gesture types is contained in the movement of the hand during the stroke phase. The stroke phase in these gestures is usually preceded by a distinct preparation phase and followed by a

retraction phase. As with spatiographic gestures, the phases can be determined by the movement and place of articulation primes. Movements up into the gesture space are preparation phases. Movements down and back from gesture space are retraction phases. Stroke phases either follow preparation phases or other stroke phases. The entire list of feature records of the stroke occurring at the reference time gives a key-frame model of the user's hand movement. This key-frame path can then be applied to the object of reference.

"Move the car like [gesture] this" is an example of a utterance that would be accompanied by a kinetographic gesture. The hand would typically move up into gesture space in a preparatory phase, demonstrate some path of motion with a stroke phase, and then retreat to rest space with a retraction phase. The stroke phase would coincide with the reference "like this." The list of hand keyframes for the stroke phase can be applied to the car to produce the animation.

Certain kinetographic gestures indicate transformations of the reference object. The movement of the hand provides a before and after image of this transformation, such as might accompany the utterance "Rotate the cup this way." Here the stroke phase can be determined as with other types of kinetographic gestures. The first feature record of the stroke can then be extracted for the before image. The last feature yields the after information.

3.2.3 How to Correlate

The timing and phase elements of the gesture determine which feature records are of interest. The next step is to determine how the hand correlates to the object. This step, which I refer to as *Iconic Mapping*, is far from trivial and will be discussed further in Section 3.3. Iconic mapping is performed by examining the hand configuration and orientation features to determine what image of the object they present. Orientation information is not always provided by the gesticulation. In some cases, a loose hand configuration will be used when only position information is being presented.

3.2.4 Transferring the Model

After the three questions outlined above have been satisfied, the model of the hand may be transferred to the computer's representation of the object of reference. In the case of a spatiographic gesture, a single record of hand location and orientation in gesture space can be used to disambiguate the position and orientation of the object referred to in speech. With kinetographic gesture, a list of hand locations and orientations can be used to key-frame the movement of an object. A list of positions and orientations can be combined to develop an image with pictographic gesture. In the special case of transformation kinetographics, two records may be used to demonstrate the before and after images of an object.

For this transfer to take place, a coordinate system transformation must exist between the user's gesture space and the graphics space of the computer. In some circumstances the orientation of one of the objects portrayed in the speaker's gesture space may reveal the orientation of the gesture space as relative to that object. In other cases, the user will adapt the orientation of their gesture space to reflect the orientation of the image being displayed on the computer monitor.

Scale must also be a part of this transformation. When showing the relative placement of two objects, for example, the vector between the hands shows direction, but distance is implied. I will suggest the following factors as yielding clues to the scale of the speaker's gesture space. When discussing objects which one might hold, the scale of gesture space is usually actual size. Objects may be placed in this space and the listener visualizes them full size. Other objects, either too large or too small to be comfortably held, are often scaled up or down to take on the size of the hand. In special circumstances, the scale may be determined by the hands themselves. If two facing sides of a box of known size are portrayed with a two-handed gesture, for example, the interpreted scale of the gesture space can be determined precisely for the given context.

3.3 Iconic Mapping

Gestures are powerful because of the images that they can present. Hands have amazing flexibility to represent objects. The area of imagistic interpretation of hand postures is largely unexplored. The interpretations of hand configurations that I discuss in this chapter are largely the result of my personal theories and observations, as well as ideas that have come out of discussions with members of my research group. It is beyond the scope of this thesis, and it would be a solid dissertation in itself, to perform the necessary observations with research subjects to determine a comprehensive set of interpretation rules.

A major factor in the mapping of iconic gesture is the determination of whether the hand configuration represents the hand acting on the object, or the object itself. The difference often depends on the point of view in the speaker's narrative. If the speaker is indicating that someone was doing something to an object, the gesture is frequently mime-like, showing how the object might be held or acted upon by that person. If the speaker is describing from an observer point of view, the hand will become the object, somehow representing the shape of that object. Some gesture classifications, such as Efron's [1941], make a distinction calling gestures acting on an object *pantomimic*, reserving the label *iconic* for gestures where the hand becomes the object. I prefer to use the term *iconic* broadly, encompassing both situations. I will adopt the descriptions *pantomimic* and *strict-iconic* to subdivide iconics only to make the distinction in the process of iconic mapping.

General rules for iconic mapping can be specified for simple objects. More complex objects may be mapped by either associating them in classes of basic objects they resemble, or by decomposing them into simpler objects, such as representing a table by it's bounding box. In the end, object shape alone is often not enough to perform iconic mapping. Image perception of objects portrayed in mime-like and strictly iconic gesture sometimes relies strongly on socially ingrained background knowledge.

3.3.1 Extracting Basic Shapes from Pantomimic Hand Configurations

We can construct mental images of three dimensional objects based on component zero, one, and two dimensional pieces. Mime-like gesture often produces images by using the hands to present points, axes, and surfaces. Significant points on an object are often represented by a grasping hand configuration. By themselves, these points lack sufficient definition to provide an iconic mapping. Using two hands opposite one another is useful for portraying a major axis of an object.

Herranz [1992] demonstrated the usefulness of identifying the principal axes of objects for manipulations such as rotation. In his demonstration, a user could indicate rotations using two-handed gesture. With objects such as airplanes, the user indicated the desired rotation by holding both hands in opposition and rotating both around some point. The users would place the hands as if they were grabbing one of the major axes, such as the wings or ends of the fuselage, and then move as if grabbing the actual airplane and rotating it.

Surfaces are easily portrayed using the flat of the hand. With the palm and fingers outstretched straight, a flat surface is easily demonstrated. Bending the fingers down 90 degrees provides a good representation of the corner of two meeting planar surfaces. Adding the thumb introduces a third surface. Curved surfaces can be portrayed easily with the bend of the fingers. Even a sphere can be easily suggested with a two-hand gesture.

Interpretation of flat surfaces and corners can be accomplished using a model of the object that specifies their location and orientation. Simple curvature of objects such as cylinders can also be specified, allowing for easy mapping of hand posture to object orientation. Complex objects would require a much more sophisticated method of determining the shape bounded by one- and two-handed gestures.

The orientation of many objects cannot be uniquely determined by merely demonstrating a major axis or surface. For manipulations of objects displayed for the user, the starting

orientation of the gesture can be correlated to the starting orientation of the object. Surfaces portrayed in gesture space may be mapped to the surfaces in virtual space that have roughly the same orientation from the user's point of view. The resulting manipulation of the hand representation can then be applied directly to the object.

3.3.2 Extracting Basic Shapes from Strict-Iconic Hand Configurations

In strictly iconic gesture, the hand becomes a representation of the object itself. As with pantomimic configurations, basic objects are often constructed out of surfaces and major axes. The flat hand becomes a representation for a great many objects, from automobiles to people to brick walls. A fist works equally well for more boxy objects. The fingers are often brought into play to demonstrate major axes or other appendages. A thumb sticking straight up from a fist can portray a bottle, for example, while stretching the index finger out suggests a gun. As with the pantomimic gestures, correspondence between the gesture and the screen representation can help to disambiguate.

3.3.3 Socially Accepted Interpretations

General iconic mapping rules for basic objects can only go so far; interpretation of many more complex objects requires previous knowledge of social implications. While the introduction of social knowledge often leads deep into the "Turing tar-pit"[2], the application of iconic mapping requires this knowledge in a very limited context.

This social knowledge plays a part in disambiguating possible interpretations of the basic iconic mappings. A flat hand, for example, can be used to demonstrate a human figure. Only cultural information, however, leads us to interpret the front of the hand as the front of the figure. Other interpretations can be traced to the construction of the human anatomy. The reason the back of the hand becomes the top of a car is most likely because

[2]Attributed to Alan Perlis, 1982, by Hutchins, Hollan & Norman [1986]

our wrists have a much broader range of motions when the palm is facing down, and this facilitates our gestural animations.

The group of objects that is perhaps the most dependent on our cultural conditioning is that of hand tools. Hand tools play an important part in our society, so it should be of no surprise that a large number of gestures reflect this. By hand tools, I mean a broad range of objects from hammers to tea-cups; from pencils to broom sticks. What is common with each of these objects is that how we hold them is grounded in our basic knowledge. Orientations of these objects can easily be specified by holding the hand in the appropriate configuration.

3.4 Combining Gesture with Speech

The gesture information can be combined in the interpreter model in a similar fashion to the linguistic information, except that the gesture channel requires the context of the speech. In a related example to the one given at the start of the chapter, a speaker says "Make a room." The model combines as before, only now the interpreter can look for spatial information in the gestures. When the noun phrase is being combined, the incoming gesture stream can be examined for a spatiographic gesture that could be applied to a "room" type object.

Two types of spatial information could be obtained here. One possibility is that a location in gesture space is indicated as the position desired for the room object. This type of reference is usually performed in relation to some point in gesture space indicated by an earlier context. The second type of information is that which sets up this type of context. When "room" is spoken, the user might hold both hands out as if holding a box. This sets up a square region that represents the room in gesture space. If the next command is "Add a couch here," another spatiographic gesture might be used to indicate a point within that square that determines the relative location of the chair.

The parsing the gesture requires several pieces of contextual information including the

time near when gesture may have occurred, the type or types of gesture being sought, the object for which iconic (i.e. hand configuration) mapping will be made, and all types of information being sought. The gesture parser can then scan the gesture segments around the desired time, examine each of these gesture segments for appropriate phase of gesture given the gesture type, attempt to map the posture and orientation to the object type given, and return the information requested (where possible). The interpreter may then use this information to fill in empty slots in the semantic model, and the task for the completed model can then be initiated.

Chapter 4

The VECIG System

The Virtual Environment Coverbal Iconic Gesture (VECIG) interface system was developed as a way to demonstrate some of the theories presented in this thesis and to provide a platform for further research of multi-modal reference. Currently, this system allows a user to interact with a simple 3-D virtual environment through the use of combined speech and gesture.

4.1 System Overview

A system block diagram for the VECIG system is shown in Figure 4-1. The system consists of the following parts.

- The Hand Input Hardware
- The Speech Recognition System
- The Gesture Processing Modules
- The Gesture Parser

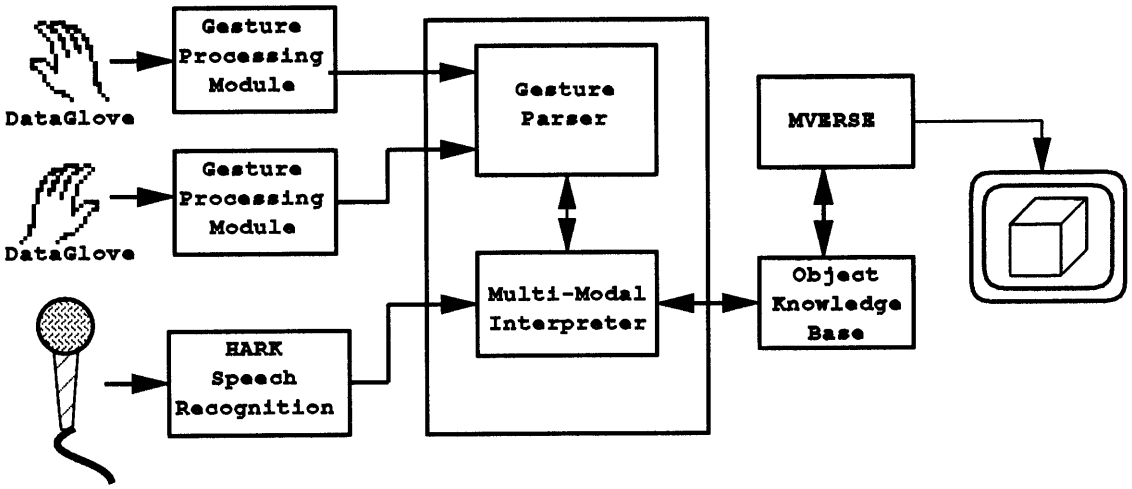

Figure 4-1: VECIG Block Diagram

- The Multi-Modal Interpreter
- The Object Knowledge Base
- The MVERSE Graphics System

The following sections provide a brief overview of each of the component parts. A detailed description of the portions specifically pertaining to this thesis is given in Section 4.2.

4.1.1 The Hand Input Hardware

One of the goals of this research was to develop a device-independent method of gesture recognition. There are several commercially available hand sensing devices on the market today, and the future quite possibly holds hand tracking by cameras at a distance. The type of hand input device used here was the VPL DataGlove Model 2 System. This input device senses the flex (extension and flexion) of the upper two joints on each finger, as well as the flex of both joints on the thumb. It does not provide any information about the spread of the fingers (abduction and adduction), or to the spread or rotation of the

thumb. Sturman [1992] provides a detailed explanation of the degrees of freedom of the hand, and details common hand input devices.

The VPL system is integrated with a Polhemus 3Space position sensing device. This device provides the X, Y, and Z coordinates of the hand in 3-space, as well as roll, pitch, and yaw. This data, along with the finger joint angles, is passed to the system at up to 30 records a second over an RS-232 line. The data is received on an HP835 workstation and time stamped by a Real Time Interface board.

4.1.2 The Speech Recognition System

The speech recognition system is a beta-test version of a soon to be released, commercially available software package developed by BBN Systems and Technologies [1], called HARK. This is a continuous, speaker independent speech recognizer that runs on a Silicon Graphics Indigo workstation. As with the gloves, the input to the interpreter should be device independence. Another requirement is that each of the words detected be time-stamped. To meet both of these requirements, a post-processor was developed to interface the HARK system to the Multi-Modal Interpreter. This module converts the word data into a compact device-independent format, and calculates a real-world time from the internal BBN timecode. It also provides error detection and is responsible for control of the BBN package.

The speech post-processor was developed by the author in C.

4.1.3 The Gesture Processing Modules

The development of the gesture processing modules was a significant part of the research for this project. This code is responsible for interfacing with the VPL DataGloves and

[1] A division of Bolt Beranek and Newman

processing the code into gesture segment, or "gestlet",[2] format for the Gesture Parser. This code also served as a stand-alone package for analyzing the hand data, which was useful for determining a suitable set of features.

During system operation, there are two separate but identical modules running, one for the left glove and one for the right glove. The modules start off by initializing the appropriate glove box and then guiding the user through glove calibration. Once calibration is complete, the module continuously reads in the raw data from the glove. Each raw data record is filtered to reduce the random error of the various values. Every record is then processed into the feature-based representation, and its features are compared against those of the most recent stored record. If the features in the new record differ from those of the previous record, the new record is extracted for further processing. These key frame records are passed to an output buffer to be gathered into gesture segments.

The output buffer gathers the features into groups that constitute a single hand movement. For the current implementation, the gestlet format consists of a record tagged as a "stop," followed by any number of records where the hand was moving, and ending with the next stop. When a complete gestlet feature list has been formed, it is output to the gesture parser. The output buffer then starts queueing up another gestlet.

Several versions of The Gesture Processing Module were developed by the author, with significant changes occurring as research progressed. The final version was written in C.

4.1.4 The Gesture Parser

The Gesture Parser is an integral part of the Multi-Modal Interpreter. This block is responsible for reading in and storing the gestlets until a request for gesture information is initiated by the interpreter. When a request is received, this module is also responsible for analyzing the gestlet queue to recognize the desired gesture, and for extracting the

[2]The term *gestlet* was inadvertently coined by the author as an abbreviated way of referring to gesture segments in code. See Koons, Sparrell and Thorisson [1993]

requested information from that gesture.

When a new gestlet is received by the parser, it is first put into a list format that is easy to process. The incoming features are put in a list, and this list is put in a structure with several high level descriptors of the movement. This structure is then put in another list with previously received gesture segments.

Hand data and speech data both have different latencies into the system, causing possible timing problems. Realignment of the gesture data takes place in the gesture parser. New requests from the Interpreter are processed immediately, examining the gestlet list for recognizable gestures. If no appropriate hand movements are found, the request is queued until more information becomes available.

Parsing of the data is accomplished by applying one or more models to the list of gesture segments. The model used is dependent on the class or classes of gesture being sought. When recognition of an appropriate gesture has been completed, a frame is created and passed back to the interpreter containing as much of the requested information as could be determined.

The Gesture Parser was implemented by the author in LISP.

4.1.5 The Multi-Modal Interpreter

The Multi-Modal Interpreter is responsible for receiving speech data from the speech recognition system, and interfacing with the Gesture Parser to collect information about the user's actions. The incoming speech elements are placed into frames and the frames are combined into phrases as applicable.

When necessary or optional information slots are empty, the interpreter will look to the other modes, here the gesture channel. The interpreter will send requests to the gesture parser for any gesture types that might be possible given the context of the speech. The Gesture Parser will then analyze the gesture data in an attempt to recognize gestures of

those types, and return information as available.

The Multi-Modal Interpreter is also responsible for the interface to the Object Knowledge Base. The interpreter can query the knowledge base for current status information about the objects instantiated, or for model information about objects in general. The interpreter also sends updates to the knowledge base as a result of the interpretation of the user's input. When construction of a complete input sequence is successful, the interpreter will send the appropriate command sequences to the knowledge base to carry out the desired action.

The Multi-Modal Interpreter is based on previous such interpreters (Thorisson, Koons & Bolt [1991] and Koons, Sparrell & Thorisson [1993]) developed by the Advanced Human Interface Group, and is the subject of continuing research. The command lexicon was specifically designed for this project. The entire interpreter was implemented in LISP.

4.1.6 The Object Knowledge Base

The Object Knowledge Base is responsible for maintaining information about objects and their relationships within the virtual environment. It contains general information about object classes, such as the major and minor axes, or other shape information that may be used to map the object to hand postures. Other records include major surfaces and default orientations which define how each object interacts with others in the virtual world. The knowledge base keeps track of specific information about instantiated objects, such as color, scale, orientation, location, and child and parent objects.

The Object Knowledge Base allows primitive relationships between objects in the virtual world. These relationships include simplified interactions representing gravity and "stiction". The simple gravity model provides that one object can be placed on another object and the knowledge base will put the object in such a position that it rests on the surface of the other object. The simple stiction model provides that objects resting on some base object will remain in place on that base object, even when it is moved.

The Multi-Modal Interpreter sends commands and queries to the Object Knowledge Base, which in turn, sends back status information. The knowledge base also interfaces to the MVERSE graphics system. As the internal model of the world is updated in the knowledge base, graphics commands are sent to the MVERSE system to reflect those changes.

The Object Knowledge Base was designed specifically for this project, under direction of the author. It was implemented in C++.

4.1.7 The MVERSE Graphics System

The MVERSE graphics system allows high level access to 3-D graphics commands. In the current implementation, it runs on top of the Hewlett Packard graphics language Starbase. The system is designed to allow loading of objects from a standard file format. All standard object and camera manipulations are supported.

The MVERSE system was developed by the Advanced Human Interface Group as a platform for this type of research.

4.1.8 The System Layout

The operation of this system requires significant processing power, so a distributed processing network is used. The processing is spread out over five workstations as described below.

The Gesture Parser and the Multi-Modal Interpreter are the heart of the system, running on a Decstation 5000/240. The hand data is read in to an HP9000/835 workstation through a Real Time Interface Board, which time stamps the incoming data. This raw data is passed through a dispatcher to separate Decstation 5000/133s, one for the right hand, one for the left. The gesture processing code passes the abstracted data to the gesture parser.

The Speech Recognition System runs on a Silicon Graphics R3000 Indigo, and the post-processed speech data is sent to the Multi-Modal Interpreter. The Object Knowledge Base runs on the master Decstation 5000/240, receiving commands from the interpreter and sending commands to MVERSE. MVERSE runs on the HP9000/835 and displays the graphics on a 1280x1024 high-resolution display.

4.2 Gesture Recognition in Practice

The method of coverbal iconic gesture recognition outlined in this document can be broken down into two distinct phases. First, the hand information is abstracted into an intermediate representation that is a feature-based description of the hand movement. Second, once the speech information has been partially evaluated, the context of the speech can be applied to the movement descriptions in a directed search for the relevant information. These two phases of operation are the respective responsibilities of the Gesture Processing Modules and the Gesture Parser.

4.2.1 The Gesture Processing Module

The demonstration system analyzes the raw data glove records into qualitative features and associated quantitative key-frames. This representation is shown in Figure 4-2. The raw data records are processed into preliminary features. These features are used to compare the record against preceding ones. If none of the record's features differ from the preceding record, the data record is destroyed. The preliminary features are then processed into the feature primes. The processed records are then combined into groups of movements representing gesture segments.

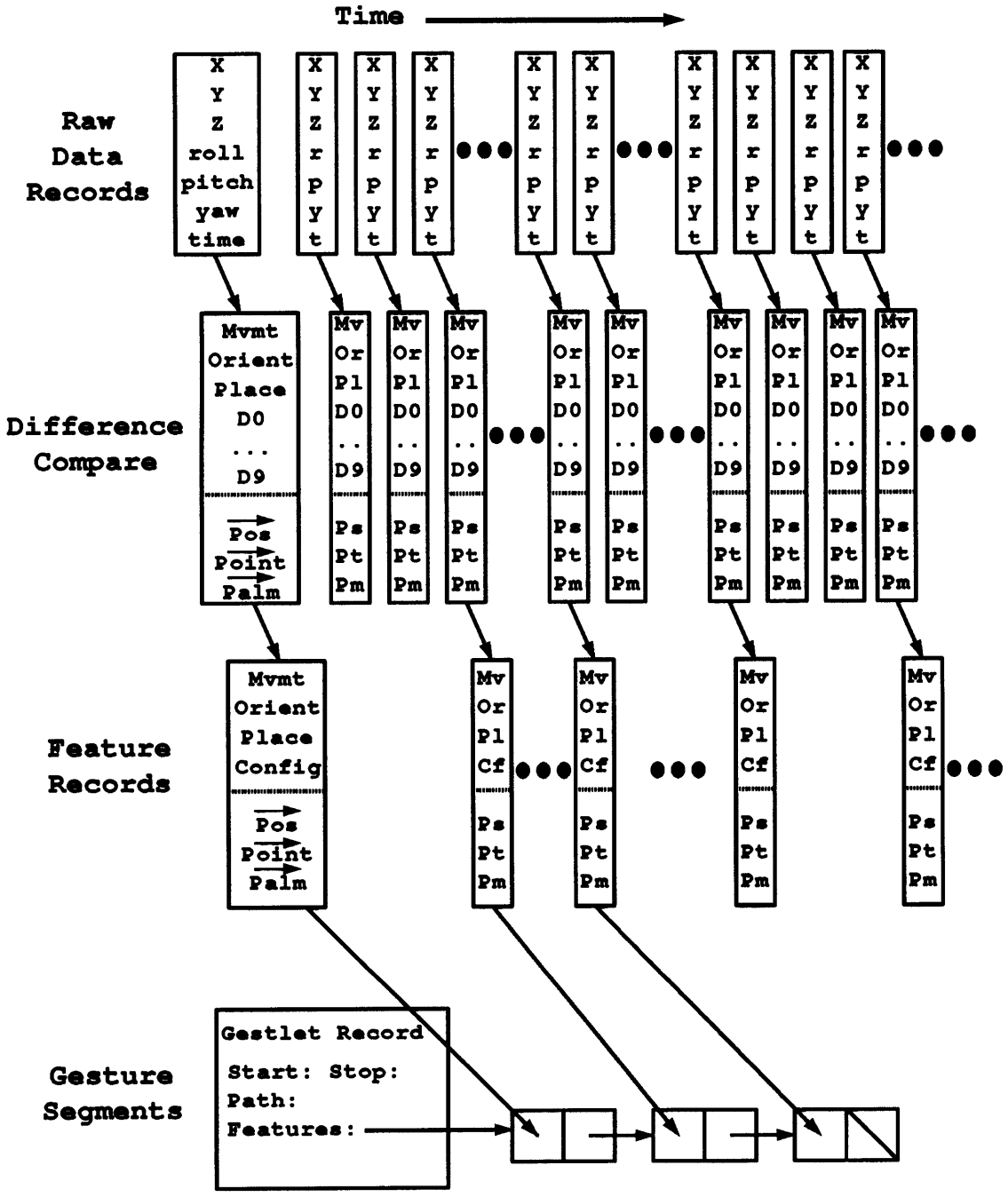

Figure 4-2: Gesture Data Flow

Determining Hand Configuration

The joint angles of the finger are quantized into preliminary descriptors of straight, relaxed or closed. The descriptors for each of the ten joints measured are then used to determine the hand configuration in a pattern matching scheme similar to fuzzy logic. An example configuration programming for "G" is shown in Figure 4-3. Each joint flex possibility is given a score from 0-10 depending on how strongly it contributes to the configuration. An incoming data record is compared against this table to determine a score representing how well the current hand shape fits that posture.

Several methods for determining hand shape have been demonstrated in previous systems (Kramer & Leifer [1989], Murakami & Taguchi [1991]). This method was chosen because it is easy to modify, and works well on top of the finger quantization. The system is able to distinguish the 11 hand configuration primes shown in Figure 4-4.

Determining Place of Articulation

The place of articulation primes are determined by a fixed "rest" volume close to the user's body. Any hand position within this volume yields a *rest* space prime, any other position results in a *gesture* space prime. An ideal system would calculate a moving gesture space based on the user's position and orientation. This cannot be done in this system due to the lack of body tracking. The exact position of the hand is maintained as part of the key-frame.

Determining Orientation

The hand orientation is determined by generating a palm vector and a pointing vector out of the hand (refer to Figure 2-2). These records are stored as part of the key-frame, and are also used to determine the orientation primes. The primes are generated by comparing the vectors to the closest axis of the fixed coordinate system of the user's gesture space,

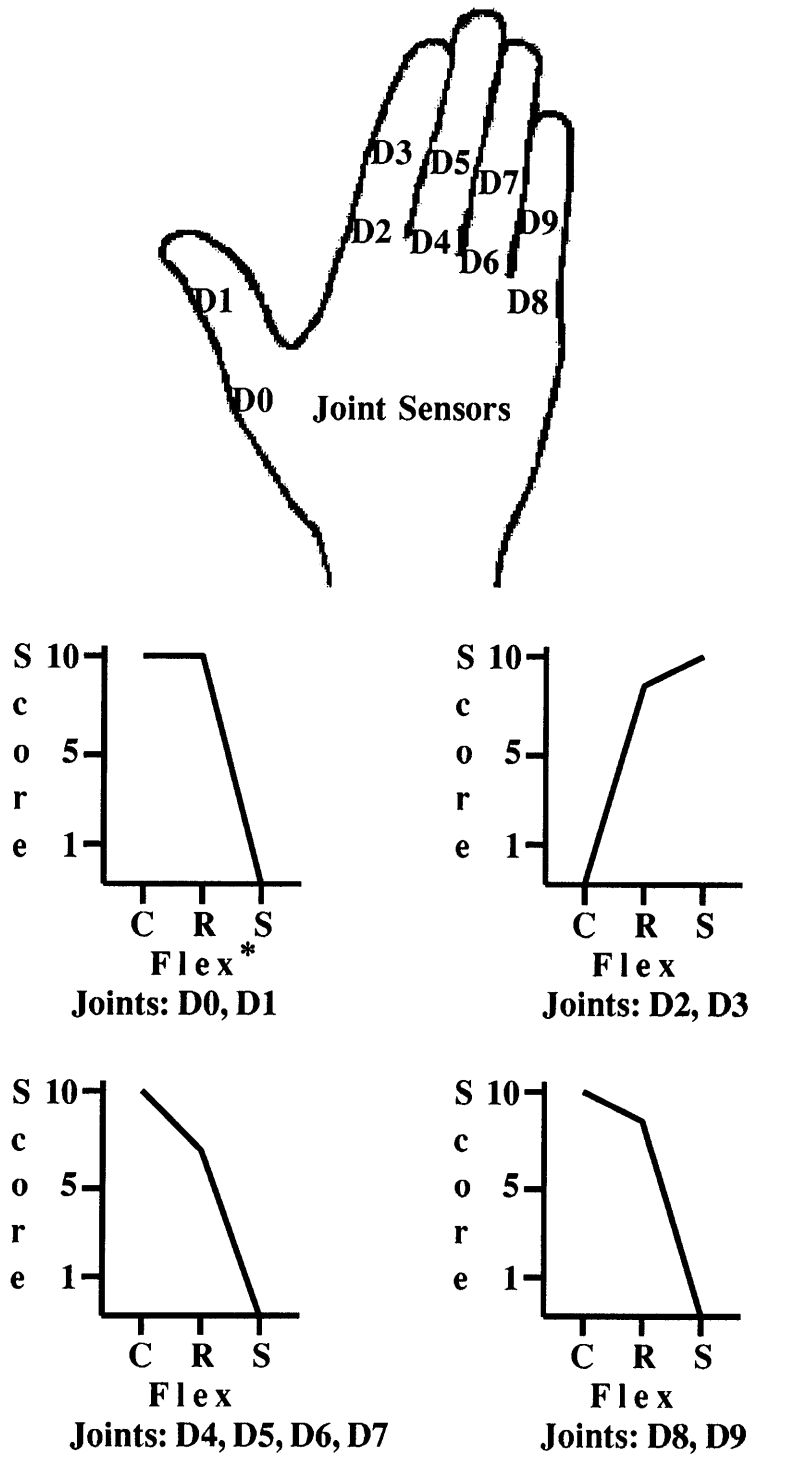

Figure 4-3: Programming of Configuration Prime "G"

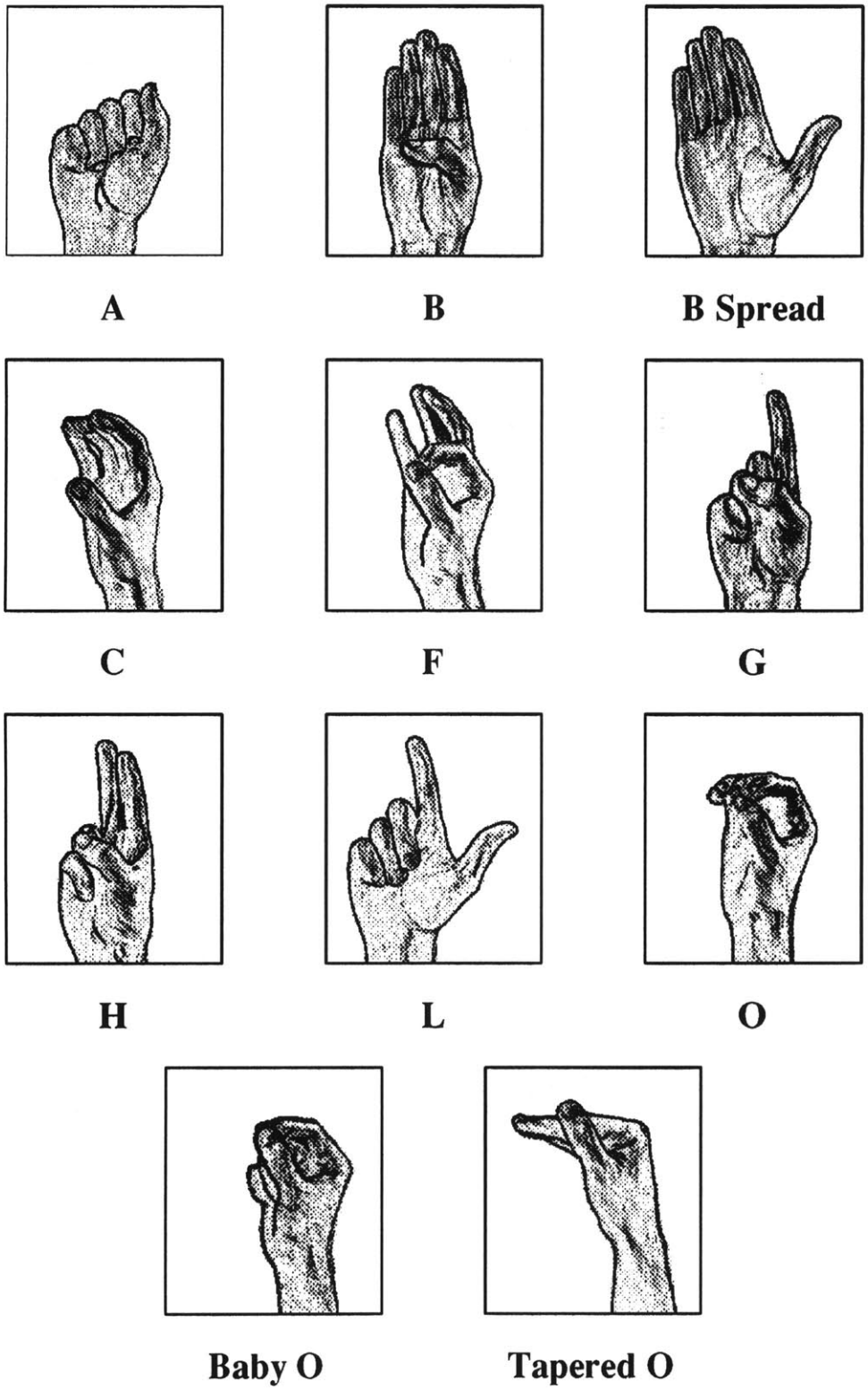

Figure 4-4: The VECIG Configuration Primes.

yielding the descriptors of up, down, left, right, forward, and back for both vectors.

Determining Hand Motion

The hand motion is determined by examining the velocity of the hand. If the velocity of the hand exceeds a minimum noise threshold, the record is logged as a "move" prime. If the speed falls below that threshold, the movement feature is determined to be a "stop". The velocity is filtered using a five point gaussian filter.

Grouping of Features into Gestlets

A simple method was employed for grouping the feature records into gesture segments or gestlets. All distinct feature records tagged as a "move" are grouped together with the preceding and following "stop" records. This method is sufficient for capturing most gesture movements. One exception to this is the case of a "finger gesture" where the hand shape changes, but the hand stays in the same place. An additional method could be added to account for these gesticulations. A completed gestlet is parsed to produce a header containing information about which hand caused the segment, what hand movement prime was carried out by the segment, and how long the segment lasted.

Determining Hand Movement primes

The hand movement primes are determined at the gestlet level. The system is capable of determining forward, back, left, right, up, and down movements. More advanced primes, such as circular motion, would require a much more complex modeling of the hand's path. The six primes mentioned were determined by examining the start and end points of the gesture segment.

Gestlet Output

After a complete gestlet has been detected, the segment header and feature list are sent to the Gesture Parser for context-dependent interpretation.

4.2.2 The Gesture Parser

The Gesture Parser maintains a list of the incoming gesture segments. When the interpreter starts to evaluate a word phase that has the possibility of accompanying gesture, it sends a query to the gesture parser. The gesture parser examines the gestlet list for appropriate motions and returns the necessary information. If no appropriate gesture information is available, the request is queued temporarily, in the event that the desired hand information is still being processed by the Gesture Processing Module. Currently, the Gesture Parser is capable of searching for spatiographic, animation kinetographic and tranformational kinetographic gestures. Pictographic gestures have not been included in the current implementation.

Spatial References

For spatial information, the parser will look for movements where the hand is being brought into place to show a location. These types of gestures are typically limited to three phases. The preparatory and stroke phases are coincident; the hand moving into place. There is a post-stroke hold of the hand in this position, and then a retraction. If the gesture is immediately followed by another using the same hand, no retraction will occur.

When looking for a gesture of this type, the parser tries to map each segment that happens near the target time to either a stroke or retraction phase. The end time for each stroke is then compared to the target time, and the closest candidates are selected. If simultaneous gesture information exists for both hands, the movements are compared to detect two-

handed gestures. The last feature in any selected segments becomes the static reference for the iconic gesture. The position, orientation, and configuration are loaded into a frame to allow iconic mapping of the associated object.

Dynamic Animations

When dynamic animation information is required, the parser looks for prominent stroke phases. Gestures of this type usually involve distinct preparatory, stroke, and retraction phases. If the gesture is preceded by another dynamic gesture made by the same hand representing the same object, the retraction phase of the first gesture and the preparatory phase of the second gesture may be dropped.

The parser looks at all segments made close to the target time and maps each movement to be a preparatory, stroke, or retraction phase. Small movements with little change can be considered noise and are given low priority. If multiple stroke phases exist for the same hand, the one closest to the reference time is selected. If segments occur simultaneously in both hands, the movements are used as a two-handed gesture. Each feature in the stroke phase is important in this type of gesture. The key frames for each feature record yield the movement information for the object being described. A frame is loaded with the entire feature list to allow mapping of the object to the string of key-frames.

Object Transformations

Object manipulations fall into a special subclass of the dynamic animations. Here a rotation or scale might be shown with the hands. With most dynamic gestures, the parser seeks to extract a prominent stroke phase. With object transformations the parser will also need to single out a stroke phase. The difference is that only the first and last features of the stroke are needed to interpret the transformation. Object manipulation gestures provide example by showing correspondence and transformation. The hand before the stroke represents the object before the manipulation, and the hand after the stroke shows

how it has been transformed. The important features for this movement are the start and end features for the stroke gestlet. These features are loaded into a frame and passed back.

4.2.3 Iconic Mapping

The last phase of gesture recognition involves the iconic mapping of the hand posture to the object being described. For the VECIG system, objects belong to simple classes, and these base classes are composed of basic parts. The cube class, for example, contains objects that are roughly cubical in shape such as a table and a chair. Cubes are composed of parts such as corners and flat surfaces. Hand shape primitives are mapped to the parts they most closely resemble. The "tapered O" prime would be mapped to a corner for a cube class, for example, while a "B spread" prime would be mapped to a surface.

Objects may belong to as many base class groups as are appropriate. The object classes currently supported are cube, cylinder, blob, major axis, and planar surface. The major axis class indicates that the object has one or two major axes that can be gripped as a handle or grasped at either end.

4.3 The Scenario

The scenario currently supported by VECIG allows a user describe the layout of a virtual room. Objects that may be placed in the room include a table, chair, couch, teapot, glass, and a dog. Objects may be placed relative to each other using spatiographic gestures. Objects may also be manipulated or animated using kinetographic gestures. No support is currently available for referencing or creating objects using pictographic gestures.

Chapter 5

Conclusion

This work has shown that recognition of Coverbal Iconic Gesture is not only possible, but useful as well. A computer interface can be developed that allows the users a range of control through the use of natural language and spontaneous gesticulation. This will allow humans to communicate with computers with the same powerful modes of communication they use with other humans.

Coverbal Iconic Gesture by its very definition is closely linked to speech. Interpretation of these gestures cannot be accomplished through a symbolic mapping to a predefined meaning; it can only be accomplished through careful consideration of how the hand movements relate to the context of speech. Recognition of this type of gesture requires a departure from previous, symbolic-based schemes. An efficient method for interpreting the gestures in the context of speech involves abstracting a feature level description of the hand movements in real-time, and extracting the essential information from these descriptions at evaluation time.

One of the goals of this thesis is to suggest a set of primitives for various aspects of the hand motion. By examining the research of others who have studied gesture in cognitive science, primitives have been suggested for the characterization of coverbal hand movements. The

primitives allow the features to be described in a form that is compact and easy to evaluate in conjunction with speech. The primitives suggested here relate to the hand configuration, orientation, movement and place of articulation and have foundations in the research of cognitive scientists studying human gesture.

Interpretation of a gesture by this method involves first determining the part of the movement that is applicable by looking at the movement and place primes together with the timing information of the accompanying speech. After a specific record or records are determined to be appropriate, the hand configuration and orientation information can be used to determine the specific location or transformation of the object being described.

The VECIG system was developed as a platform to test out these theories. This system allows a user to interact with a simple virtual environment through the use of speech and iconic gesture. A simple scenario of arranging furniture in a room gives the user a specific context in which to manipulate objects. The purpose of this system was to demonstrate some of the theories presented in this thesis, and to show a simple example of how coverbal iconic gesture recognition might be used.

5.1 Applications

The scenario for the VECIG system was chosen for its similarity to two previous systems developed in the lab. The first was the Divadlo desktop theater system developed by Steve Strassman [1991]. This system allowed automatic generation of animations that expressed emotions and intentions of synthetic actors. These actors could be directed using natural language keyboard input.

The second platform that inspired my scenario was the CINEMA system developed by Steven Drucker and Tinsley Galyean [1992]. CINEMA allowed procedural camera movements through a virtual movie set with a pre-defined animation. The user could specify such parameters as position, orientation and field of view relative to objects, events, and the general state of the environment. The purpose of CINEMA was to provide a platform

for teaching cinematographic camera and editing techniques and to demonstrate the potential of 3-D computer graphics for planning out live action film shoots. Commands to the CINEMA system were entered in a command-line format using a keyboard.

These two systems, together with VECIG, demonstrate important pieces of what could be a powerful creative environment for desktop theater. Imagine a system with semi-autonomous synthetic actors capable of expressing intentions and emotions and responding to voice and gesture input. Camera positions and movements could also be specified in much the same way that animations of other objects are specified. The results would provide a reasonable desktop theater with a simple yet powerful method of directing.

Professional CAD systems could also benefit from the addition of coverbal iconic gesture. While detailed design modifications would be better controlled through direct manipulation, overall layout and view control would be better controlled by speech and gesture.

Virtual manifestations of human descriptions could also be used to overcome the Gulf of Evaluation between humans. In our intra-human communications, we rely heavily on our ability to span the Gulf of Execution. If I were describing an accident, for example, I would trust the perceptive powers of the listener to understand my narrative. With computers we rely on feedback to verify that a command was understood. In some situations it is imperative that the listener understand. The jurors of a trial should understand the description of a witness completely. To facilitate this, models or diagrams are often constructed. Recently, 3-D animations have been used for this purpose. A system which allowed a computer illiterate individual to construct a rough animation could be of great benefit.

5.2 Future Work

This is an area where little work has been done in the past. Understandably, there are many areas where future research is needed. Three research areas in particular would greatly extend the the usefulness of this thesis.

One improvement would be to combine this work with other gesture types. Deictic gestures in particular have been shown to be of great benefit in multi-modal interactions. Previous interfaces have relied on strict pointing symbol hand configurations to signal a deictic reference. In related research, I have found that the gesture recognition scheme outlined in this thesis works well at detecting a wide range of deictic references, from pointing gestures to sweeping references to grab type motions used to indicate groups. Further work needs to be done on introducing an expanded set of deictic gestures and on determining how to disambiguate deictic from iconic gestures when the context of the speech would support either.

A second big improvement could come from an in-depth cognitive study of how we map objects to hand configurations in iconic gesture. McNeill's work falls short of developing any rules that might be followed in such interpretations. I have set forth several theories, but backing them up through experimentation with subjects is out of the scope of this work.

The third area where further work could yield improvement is in the analysis of complex hand movements. The method used in the VECIG system performs well at detecting distinct hand movements that roughly follow straight lines. Complex movements, such as circles and spirals cannot be detected as such. A better hand tracking model would allow the system to detect such motions, and possibly break some up into simpler segments. The motions detected could also be described in somewhat finer detail, revealing whether a gesture was quick and assertive or slow and meandering. The addition of such information to gesture recognition is not unlike including intonation and inflection with speech recognition and could supply some valuable clues into the characteristic motion of the object being portrayed.

In general I would expect to see a growing interest in Natural Language and Multi-modal interfaces. The time has just arrived when continuous, speaker-independent speech recognizers are becoming available. Hand-tracking hardware is getting better, and progress is being made towards tracking with cameras at a distance. Better input technology alone is not the solution, but will fuel the interest.

Bibliography

[1] Bolt, R. A. (1980). "Put-That-There": Voice and Gesture at the Graphics Interface. *Proceedings of SIGGRAPH '80*, 262-270. ACM Press, New York.

[2] Bolt, R. A. (1984). *The Human Interface.* Van Nostrand Reinhold, New York.

[3] Bolt, R. A. and Herranz, E. J. (1992). Two-handed Gesture with Speech in Multi-Modal Natural Dialogue. *Proceedings of UIST '92*, 7-12. ACM Press, New York.

[4] Butterworth, J., Davidson, A., Hench, S. and Olano, T. M. (1992) 3DM: A Three Dimensional Modeler Using a Head-Mounted Display. *Proceedings 1992 Symposium on Interactive 3D Graphics*, 135-138. ACM Press, New York.

[5] Cohen, P. R., Sullivan, J. W., et. al. (1989). Synergistic Use of Direct Manipulation and Natural Language. *CHI '89 Proceedings*, 227-233. ACM Press, New York.

[6] Drucker, S. M., Galyean, T. A., and Zeltzer, D. (1992). CINEMA: A System for Procedural Camera Movements. *Proceedings 1992 Symposium on Interactive 3D Graphics*, 67-70. ACM Press, New York.

[7] Efron, D. (1941). *Gesture and Environments.* King's Crown Press, Morningside Heights, New York.

[8] Fisher, S., McGreevy, M., Humphries, J. and Robinett, W. (1986). Virtual Environment Display System. *Proceedings 1986 Workshop on Interactive 3D Graphics*, 135-138. ACM Press, New York.

[9] Graham, J. A. and Argyle, M. (1975). The Communication of Extra-verbal Meaning in Gestures. *International Journal of Psychology.*

[10] Graham, J. A. and Heywood, S. (1976). The Effects of Elimination of Hand Gestures and of Verbal Codability on Speech Performance. *European Journal of Social Psychology*, 189-195.

[11] Hauptman, A. G. (1989). Speech and Gestures for Graphic Image Manipulation. *CHI '89 Proceedings*, 241-245. ACM Press, New York.

[12] Herranz, E. J. (1992). Giving Directions to Computers via Two-handed Gesture, Speech and Gaze. S.M. Thesis, MIT Media, Arts and Sciences Section.

[13] Hutchins, E. L., Hollin, J. D., and Norman, D. A. (1986) Direct Manipulation Interfaces. In *User Centered Systems Design*, D. A. Norman and S. W. Drayer, ed., 87-124. Lawrence Erlbaum Associates, Hillsday, N.J.

[14] Kendon, A. (1980). Gesticulations and Speech: Two Aspects of the Process of Utterance. In *The Relation between Verbal and Non-verbal Communication*, M. R. Key, ed., 207-227. Mouton, The Hauge.

[15] Klima, E. and Bellugi, U. (1979). *Signs of Language.* Harvard University Press, Cambridge, MA.

[16] Koons, D. B., Sparrell, C. J., and Thorisson, K. R. (1993). Integrating Simultaneous Input from Speech, Gaze, and Hand Gestures. To be published in *Intelligent Multi-Media Interfaces*, M. Maybury, ed. AAAI Press, Menlo Park, CA.

[17] Koons, D. B. (1993). Capturing and Interpreting Multi-Modal Descriptions with Multiple Representation. In preparation.

[18] Kramer, J. and Leifer, L. (1989). The Talking Glove: An Expressive and Receptive "Verbal" Communication Aid for the Deaf, Deaf-Blind, and Nonvocal. Department of Electrical Engineering , Stanford University, Stanford, CA.

[19] Laurel, B. K. (1986). Toward the Design of a Computer-Based Interactive Fantasy System. Ph.D. Dissertation, Ohio State University.

[20] Laurel, B. K., ed., (1990). *The Art of Computer Interface Design.* Addison-Wesley Publishing Company, Reading, MA.

[21] McNeill, D. and Levy, E. (1982). Conceptual Representations in Language Activity and Gesture. In *Speech, Place, and Action*, Jarvella, R. J. and Klein, W., ed., 271-295, John Wiley & Sons Ltd.

[22] McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought.* The University of Chicago Press, Chicago.

[23] Murakami, K. and Taguchi, H. (1991). Gesture Recognition Using Recurrent Neural Networks. *CHI '91 Conference Proceedings*, 237-242. ACM Press, New York.

[24] Neal, J.G. and Shapiro, S.C. (1991). Intelligent Multi-Media Interface Technology. In *Intelligent User Interfaces*, J.W. Sullivan and S.W. Tyler, eds., 11-43. ACM Press, New York.

[25] Nespoulous, J., Perron, J. and Lecours, A. R. (1986). *Biological Foundations of Gestures: Motor and Semiotic Aspects.* Lawrence Erlbaum Associates, Hillsdale, New Jersey.

[26] Poizner, H., Klima, E. S. and Bellugi, U. (1987). *What the Hands Reveal about the Brain.* MIT Press, Cambridge, MA.

[27] Rime, B. and Sciaratura, L. (1991). Gesture and Speech. In *Fundamentals of Nonverbal Behavior*, R. S. Feldman and B. Rime (ed.), 239-281. Press Syndicate of the University of Cambridge, New York.

[28] Rubine, D. (1991). The Automatic Recognition of Gestures. Ph.D. Dissertation, Carnegie-Mellon University.

[29] Stokoe, W. C. (1960). Sign Language Structure: An Outline of the Visual Communication System of the American Deaf. *Studies in Linguistics*, Occasional papers No. 8. University of Buffalo Press, Buffalo, NY.

[30] Stokoe, W. C. (1972). *Semiotics and Human Sign Languages.* Mouton, The Hauge.

[31] Strassmann, S. H. (1991). Desktop Theater: Automatic Generation of Expressive Animation. Ph.D. Dissertation, MIT Media, Arts and Sciences Section.

[32] Sturman, D. J. (1992). Whole-hand Input. Ph.D. Dissertation, MIT Media, Arts and Sciences Section.

[33] Thorisson, K. R., Koons, D. B. and Bolt, R. A. (1992). Multi-Modal Natural Dialogue. *CHI Proceedings '92*, 653-654. ACM Press, New York.

[34] Thorisson, K. R. (1993). Dialogue Control in Social Interface Agents. *INTERCHI Adjunct Proceedings '93*, 139-140. ACM Press, New York.

[35] Vere, S. A., (1991). Organization of the Basic Agent. *SIGART Bulletin*, 164-168. ACM Press, New York.

[36] Weimer, D. and Ganapathy, S. K. (1989). A Synthetic Visual Environment with Hand Gesturing and Voice Input. *CHI '89 Proceedings*, 235-240. ACM Press, New York.

[37] Whittaker, S. and Walker, M. A. (1991). Toward a Theory of Multi-Modal Interaction. *AAAI '91 Workshop Notes*. AAAI, Press, Menlo Park, CA.

Appendix A

Examples of Interaction

This appendix outlines several excerpts representative of the VECIG scenario interaction. Three examples are given. In the figure for each example, the reference word that triggers the search for gesture information is underlined. A transcript of word and gesture input is given at the end of this section.

The first example is shown in Figure A-1. This is a spatiographic gesture where the speaker first sets up the context of the gesture space. Both hands are brought into gesture space in the "tapered O" shape representing the far corners of the table. By correlating the hand positions with the coordinates of the table, a transformation between gesture space and object space is determined. After the "on the table" context has been setup, the left hand is moved to refer to a position within this context. The point at which the hand comes to rest is then used to determine a relative position for the teapot on the table. The resulting action is shown in Figure A-2.

The second example (Figure A-3) shows another spatiographic gesture. Here the location of a new object (the glass) is being shown in relation to an existing object (the teapot). The left hand comes to rest in gesture space about the time the word "glass" is spoken, with an "O" hand shape representing the glass. The right hand joins the left in gesture

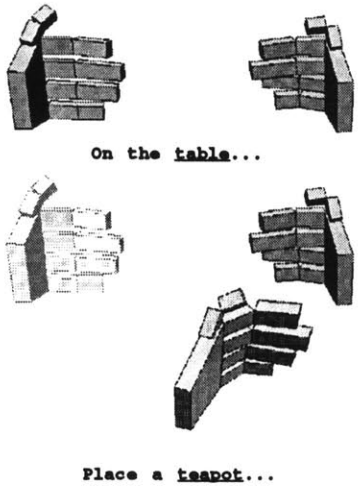

Figure A-1: Example of specifying context for spatiographic placement.

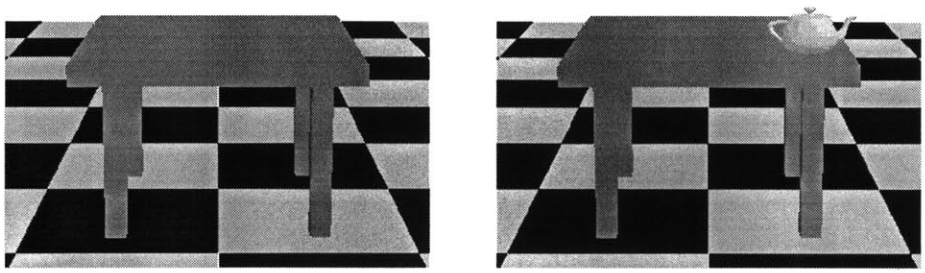

Figure A-2: Placement of the teapot on the table.

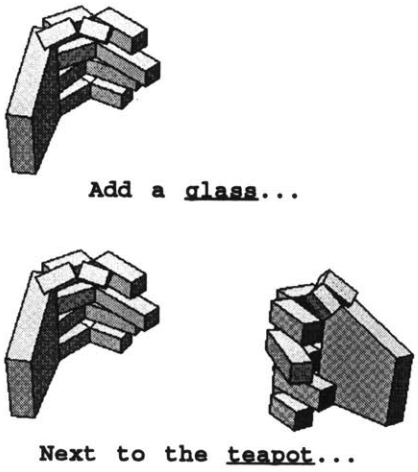

Figure A-3: Example of relative position spatiographic placement.

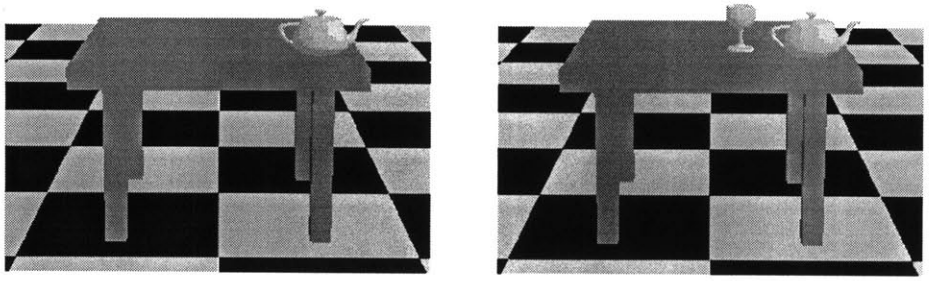

Figure A-4: Placement of the glass next to the teapot.

space with an "A" shape about the time the word "teapot" is spoken. By comparing the positions where the two hands come to rest, the relative direction of the glass is determined. Scaling information is determined by looking at what context the two objects share. Here, both the glass and the teapot are on the table, and the table has been recently established in gesture space, so the scale is assumed to be the same as before. The "O" shape portrays the hand wrapped around a cylinder. The cylinder configuration is a good match to the glass and vertical orientation (which also happens to be the default) is extracted from the vertical cylinder of the hand. The system response is shown in Figure A-4.

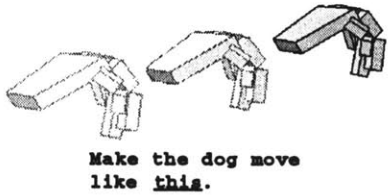

Figure A-5: Example of kinetographic gesture.

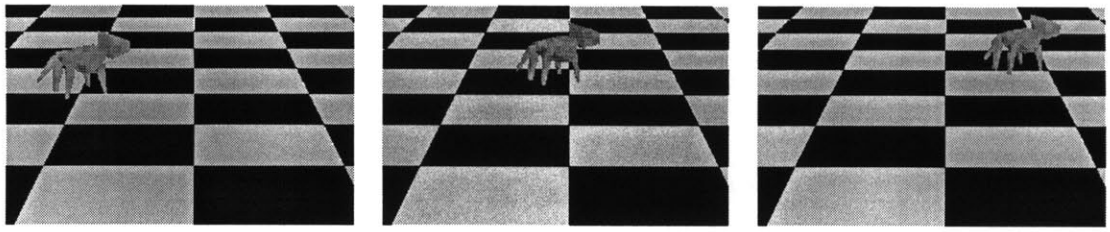

Figure A-6: Still-frames of dog animation.

The final example (Figure A-5) demonstrates a kinetographic gesture referring to a path. The speech interpreter looks for gesture information when the reference "like this" occurs. The left hand moves up into gesture space in anticipation of the stroke phase. The stroke phase indicates the desired path of the dog and terminates right around the time the word "this" is spoken. The feature list from this stroke is used to generate a smooth animation for the dog along the desired path. Three still-frames representing this action are displayed in Figure A-6.

Parser Input Transcription

The following is a transcript of word and gestlet input into the multi-modal interpreter. All time codes are in hundredths of seconds. Only the first and last features of each gestlet have been printed out for brevity.

```
Key:
Gestlet <hand> Strt: <start-time> Stp: <stop-time> Feats: <# features>
Feature: @ <time> Mtn: <stop or start> Place <rest or gest space>
   <x y z>
   H: <Hand shape> Pt:  <point direction> <point vector x y z>
                   Plm: <palm  direction> <palm vector x y z>
****
Gestlet left : Strt: 416688 Stp: 416822 Feats: 8 Mv: 2 dl:  15.52
  Feature: @ 416688  Mtn: S Plc: R (  4.50, -4.96,-21.28)
   H:O  Pt: F (  0.16,  0.31,  0.94) Pm: R (  0.95, -0.31, -0.05)
  Feature: @ 416822  Mtn: S Plc: G ( -4.32,  1.02, -9.99)
   H:tO Pt: F ( -0.26,  0.15,  0.95) Pm: R (  0.96,  0.17,  0.24)
****
Gestlet right: Strt: 416688 Stp: 416815 Feats: 7 Mv: 3 dl:  11.56
  Feature: @ 416688  Mtn: S Plc: R ( 13.84, -4.77,-16.02)
   H:L  Pt: L ( -0.64,  0.42,  0.64) Pm: L ( -0.71, -0.65, -0.28)
  Feature: @ 416815  Mtn: S Plc: G ( 12.79,  1.91, -6.64)
   H:tO Pt: F ( -0.36,  0.05,  0.93) Pm: L ( -0.93,  0.05, -0.37)
****
Word uttered: ON! @ 416782
Word uttered: THE! @ 416788
Word uttered: TABLE! @ 416835
****
Gestlet left : Strt: 416993 Stp: 417106 Feats: 7 Mv: -2 dl:  12.59
  Feature: @ 416993  Mtn: S Plc: G ( -4.07,  0.79, -9.63)
   H:bO Pt: F ( -0.24,  0.13,  0.96) Pm: R (  0.95,  0.25,  0.20)
  Feature: @ 417106  Mtn: S Plc: G (  8.30,  0.52,-11.94)
   H:C  Pt: F (  0.50,  0.34,  0.80) Pm: R (  0.85, -0.38, -0.37)
****
Word uttered: PLACE! @ 417024
Word uttered: A! @ 417029
Word uttered: TEAPOT! @ 417083
****
Gestlet left : Strt: 417184 Stp: 417340 Feats: 10 Mv: -3 dl:  13.35
  Feature: @ 417184  Mtn: S Plc: G (  8.26,  0.46,-11.86)
   H:C  Pt: F (  0.48,  0.33,  0.81) Pm: R (  0.86, -0.36, -0.36)
  Feature: @ 417340  Mtn: S Plc: R (  4.45, -5.59,-23.13)
   H:bO Pt: F (  0.48,  0.40,  0.78) Pm: R (  0.88, -0.25, -0.41)
****
Gestlet right: Strt: 417207 Stp: 417345 Feats: 8 Mv: -3 dl:  14.10
  Feature: @ 417207  Mtn: S Plc: G ( 13.02,  1.98, -6.77)
   H:tO Pt: F ( -0.37,  0.04,  0.93) Pm: L ( -0.93,  0.05, -0.38)
  Feature: @ 417345  Mtn: S Plc: R ( 14.85, -6.64,-17.78)
   H:L  Pt: L ( -0.83,  0.15,  0.54) Pm: B ( -0.56, -0.34, -0.76)
****
Gestlet left : Strt: 417996 Stp: 418100 Feats: 7 Mv: 3 dl:  13.13
  Feature: @ 417996  Mtn: S Plc: R (  4.08, -4.39,-22.57)
   H:bO Pt: F (  0.46,  0.29,  0.84) Pm: R (  0.89, -0.12, -0.44)
  Feature: @ 418100  Mtn: S Plc: G (  2.59,  0.60,-10.52)
   H:O  Pt: F ( -0.06,  0.32,  0.94) Pm: R (  0.99, -0.08,  0.09)
****
Word uttered: ADD! @ 418009
Word uttered: A! @ 418014
Word uttered: GLASS! @ 418047
****
```

```
Gestlet right: Strt: 418196 Stp: 418298 Feats: 7 Mv: 3 dl:  12.24
  Feature: @ 418196  Mtn: S Plc: R ( 14.96, -6.63,-17.55)
   H:O  Pt: L ( -0.80,  0.10,  0.59) Pm: B ( -0.60, -0.21, -0.78)
  Feature: @ 418298  Mtn: S Plc: G ( 12.08,  0.53, -8.05)
   H:A  Pt: L ( -0.67,  0.31,  0.67) Pm: L ( -0.73, -0.40, -0.55)
****
Word uttered: NEXT! @ 418216
Word uttered: TO! @ 418226
Word uttered: THE! @ 418236
Word uttered: TEAPOT! @ 418286
****
Gestlet right: Strt: 418368 Stp: 418517 Feats: 9 Mv: -3 dl:  11.74
  Feature: @ 418368  Mtn: S Plc: G ( 12.03,  0.57, -7.99)
   H:A  Pt: L ( -0.68,  0.27,  0.68) Pm: L ( -0.72, -0.41, -0.56)
  Feature: @ 418517  Mtn: S Plc: R ( 14.42, -6.36,-17.16)
   H:O  Pt: L ( -0.83,  0.14,  0.54) Pm: B ( -0.55, -0.35, -0.75)
****
Gestlet left : Strt: 418379 Stp: 418543 Feats: 10 Mv: -3 dl:  13.33
  Feature: @ 418379  Mtn: S Plc: G (  2.32,  0.93,-10.88)
   H:tO Pt: F ( -0.04,  0.35,  0.94) Pm: R (  1.00, -0.01,  0.05)
  Feature: @ 418543  Mtn: S Plc: R (  4.46, -5.63,-22.28)
   H:bO Pt: F (  0.34,  0.33,  0.88) Pm: R (  0.93, -0.22, -0.28)
****
Gestlet left : Strt: 491776 Stp: 491888 Feats: 7 Mv: 2 dl:  13.84
  Feature: @ 491776  Mtn: S Plc: R (  5.21, -5.11,-22.15)
   H:O  Pt: F ( -0.09,  0.48,  0.87) Pm: R (  0.85, -0.42,  0.32)
  Feature: @ 491888  Mtn: S Plc: G ( -1.48,  6.53,-18.78)
   H:O  Pt: F (  0.01,  0.59,  0.81) Pm: R (  0.57, -0.67,  0.48)
****
Word uttered: MAKE! @ 491866
Word uttered: THE! @ 491874
Word uttered: DOG! @ 491900
Word uttered: MOVE! @ 491924
****
Gestlet left : Strt: 491921 Stp: 492053 Feats: 8 Mv: 2 dl:  17.19
  Feature: @ 491921  Mtn: S Plc: G ( -1.77,  6.13,-18.12)
   H:O  Pt: F (  0.05,  0.58,  0.81) Pm: D (  0.51, -0.71,  0.49)
  Feature: @ 492053  Mtn: S Plc: G (  8.46,  5.21, -4.33)
   H:C  Pt: F (  0.07,  0.26,  0.96) Pm: D (  0.10, -0.96,  0.25)
****
Word uttered: LIKE! @ 491969
Word uttered: THIS! @ 492012
****
Gestlet left : Strt: 492072 Stp: 492238 Feats: 10 Mv: -3 dl:  20.31
  Feature: @ 492072  Mtn: S Plc: G (  8.44,  5.30, -4.35)
   H:C  Pt: F (  0.07,  0.26,  0.96) Pm: D (  0.10, -0.96,  0.25)
  Feature: @ 492238  Mtn: S Plc: R (  5.40, -4.68,-21.77)
   H:O  Pt: F (  0.00,  0.52,  0.85) Pm: R (  0.61, -0.68,  0.42)
****
```