Gesture Analysis for Graphic Manipulation

Alexander G. Hauptmann, Paul McAvinney and Sharon R. Shepard
28 November 1988
Technical Report CMU-CS-88-198

This research was sponsored in part by the National Science Foundation SBIR grant No. ISI-8660349.

Table of Contents

List of Figures

List of Tables

1 Abstract

The primary objective of this research was to evaluate the efficacy of gestures to manipulate graphic objects. The knowledge about how people intuitively use gestures to communicate with computers will be beneficial in future development of gesture based input devices.

The work performed and the results obtained in this report include:

- Evaluated the efficacy of gestures: A simulated graphic manipulated computer workstation was developed that included video and voice recorders. Three populations were tested. These included computer literate, computer naive and spatially oriented groups. All three groups were able to quickly and intuitivley use gestures and speech to describe the computer task they wanted to accomplish without hints or prompting. The task was specific, but the input directions were only to assume that the computer could see and hear the subject and to use gestures or speech to accomplish the task.
- Demonstrated a commonality of intuitive gesture and speech patterns among the tested populations: There was a large degree of commonality of gesture and speech patterns used by subjects in this experiment. This result suggests that a gesture/speech language based on common gestures and speech patterns for graphic manipulation could be intuitively used by very large numbers of people. Development of such a language would have great research and commercial value and would help make possible broad bandwidth human to computer communication.
- Developed a gesture classification scheme: A method was developed to describe and classify gestures used on review of the video tapes of subjects in this experiment. It was not only useful in the analysis of data, but further suggests that common software modules may be developed for the development of graphic manipulation systems.
- Determined a user preference for combined gesture and speech communication: During the exit interviews, subjects were asked whether they perferred gestures, speech, or the combination. There was a strong preference for both gestures and speech.

The efficacy of gestures and the commonality of gesture and speech patterns demonstrated by these experiments suggest it would be of significant value to develop a language based on the common intuitive elements. The results of this research can be applied to develop a gesture-based or gesture and speech based system which enables computer users to manipulate graphic objects using easily learned intuitive gestures to perform graphic tasks such as assembling a pump or motor, controlling vehicles or equipment, or changing the dimensions of a drawing while using an integrated graphic text editor.

2 Introduction

Recent years have witnessed an explosion in both the number and capabilities of various devices whose purpose is to facilitate communication between people and machines. In the area of computer graphics, new applications have become possible due to increasing resolution and functionality, and decreasing costs of such devices. In the realm of graphics, however, there still remains an asymmetry between the computer's ability to generate complex graphic output and it's ability to recognize and interpret similarly complex input.

The most widely used input device is the keyboard. Given the almost universal need for a device which can capture text strings quickly and reliably, it is unlikely that the keyboard will be replaced in the near future, even by speech-recognition devices. Nevertheless, using a keyboard effectively requires extensive training and practice; this is a highly unwelcome "feature" to those who need to use computers, but who are not typists. Lack of standardization in keyboard layout makes things worse. Most important in the newly-emerging context of cheap, high-resolution graphics is the fact that the keyboard does not permit simple, intuitive manipulation of graphic objects by the computer user.

A significant improvement over the keyboard in dealing with graphic objects is the mouse, a device which can be rolled on a hard surface in order to point at objects displayed on a video monitor. It is one of several devices available for pointing at displayed objects. Its most serious disadvantage is the fact that it introduces a third focus of attention for the user (in addition to the video display and the keyboard), slowing down interaction seriously in those cases where it is not practical or possible to allocate one hand continuously to holding the mouse.

An even more convenient device, at first glance, is the touch screen. This is a device which fits over the front of a computer display and senses a single finger. One of its advantages is that it does not take up desk space, a problem familiar to mouse users. Unfortunately, low-resolution touch screens tend to be of little use for anything other than menus, and high-resolution touch screens tend to be expensive. Both force the user to revert to the keyboard if it is desired to do something to the displayed object other than point at it. [19] This consideration is more important than it might at first appear. In a graphics-oriented environment we find that our usual objective is to *construct and/or manipulate* graphic objects, not just point at them. The types of computer systems that require only pointing are generally those that are restricted to menus, which force a somewhat passive role on the user. That is, the user can only select from a predetermined list of actions, rather than construct and/or manipulate new objects directly. If much of the future interaction between humans and graphics-capable computers will involve manipulating graphic objects, it makes sense to develop an input device that lets us do just that in the most direct and intuitive way, with minimal user training. This is the rationale that has driven the development of several graphic manipulators including the *Sensor Frame*, the *Rolky* [13], the *Drawing Prism* [10], and the "glove". Unlike the touch screen, mice and other devices limited to pointing (using single fingers), these devices have the additional capability for manipulating graphic objects using multiple fingers.

Given the need to construct and manipulate graphic objects, our objective was to evaluate the *efficacy* of alternative gestures in manipulating graphic objects.[1] This knowledge provides information needed to identify those classes of gestures most appropriate for object manipulation in a particular environment. The use of carefully-chosen gestures minimizes the difficulties users experience in interacting with graphic systems, and reduces the amount of information that must be remembered by users in order to perform graphic-object manipulation.

At this point, it might be useful to elaborate a bit more on what is meant by graphic object manipulation using gestures.

In the case of only one finger, if we recognize only the finger's angle in addition to its x-y coordinates, we can use knowledge of the angle to "push" objects around the screen. Two or more fingers can be used to delimit, or establish the "scope" of a subsequent gesture. For example, two fingers can establish the "scope" of text one wants to cut and paste in a text-editing environment.

---

[1] *Gestures* as used in this proposal refer to hand motions used to manipulate graphic objects.

With suitable gestures based on recognition of multiple fingers and their motions, we can move, rotate, and scale objects, or even the viewpoint of the user, in three dimensions. These examples suggested the possibility of a "gesticulative language" based on concatenated gestures, wherein the "scoping" mentioned in the above example is analogous to a noun phrase that describes "how much" text is to be acted upon by a subsequent gesture, which serves as a verb (for example *move, copy, or delete*). [14] It might also be possible to use gestures in combination with speech. Using a device capable of recognizing only 20 or 30 words of disconnected speech, we can use gestures for scoping, and speech for specifying operations. For example, we can delineate some subset of similar (or dissimilar) objects on the screen using a gesture, and say, at the same time, "disappear", or "turn red".

This latter type of operation may prove to be faster than either speech or gestures alone, and it is far more intuitive. This experiment also determined that speech and gestures combined are preferred by users. Intuitiveness is important where mistakes are expensive, because it renders mistakes less likely.

Clearly, the use of gestures provides much more expressive power than pointing. Just as a graphics terminal greatly enhances the bandwidth of communication from the computer to the user, a good graphic manipulator will enhance the bandwidth in the reverse direction, from user to computer. This makes sense when we consider that humans presently have much more knowledge in usable form than do computers. We hope to rectify this imbalance somewhat with gesture-based systems. Our objective for this project is to obtain knowledge about how users might best communicate with graphic systems. This information can then be used to design a gesticular language for graphic manipulators of the type described in the next section.

3 Background

From a detailed analysis of subject behavior, we hoped to discern general patterns and properties (if there were any) of speech and gestures used to manipulate objects. In addition, we hoped to relate classes of observed gestures to semantics.

- Gestures are a largely unexplored way of interacting with a computer. The research carried out in this report provides a first insight into the way people use gestures when interacting with computer systems. One result of the research is a partial inventory of gestures people use when communicating with a computer.
- The term *gestures* is used in this report to refer to hand motions used to manipulate graphic objects. It does not refer to hand motions as a complete substitute for verbal communication, in the sense of American Sign Language [6]. Nor does the term refer to body motions, including hands as a means of emphasis in conversation. It is important to take into account our specialized use of the term when surveying other related research on "gestures."
- In this work we attempted to answer several questions. We tested random samples of users from different backgrounds (computer-naive, computer-experienced and computer-naive but skilled in object manipulation). No hints or constraints regarding the best method of manipulating an object were given. Is there any commonality in the way naive users might attempt to manipulate a graphic object?
- Are the types of gestures and speech used for graphic object manipulation correlated with the level of computer literacy of a given population?
- What combinations of gestures and speech are used by each population? Results from this question suggests some productive work to be done in the future.
- What difficulties do users face in manipulating graphic objects? This will also be expanded upon in the future when a study of problems associated with commercially available pointing devices is examined.

The experiment followed the lines of [11].

3.1 Outside Related Research

The most relevant outside research related to this research has been done with speech. Furnas et al. [8] presented subjects with a set of operations for text-editing, want-ads, and recipes while asking them to label or name the operation. Their objective was to establish a widely shared, intuitive verbal set of commands/descriptions for these manipulations and objects. The goal was to use this as a basis for a command language vocabulary. Their result, in the verbal domain, showed that approximately 10% to 20% of the words spontaneously used by the subjects were shared. This percentage is higher under certain assumptions for synonyms.

Chapanis [2] and his group used cooperative problem solving tasks to investigate the effects of communication mode on problem-solving ability. Even though they provided communications channels similar to those proposed in our experiment, there was no analysis of the gestures exhibited by the subjects. Furthermore, communication with a computer is significantly different from interpersonal communication [19, 11].

All the empirical research related to manipulating graphic objects on a display has been restricted to pointing devices and keyboard commands [19, 15]. The severe limitations inherent in this type of communication (pointing) make the results largely irrelevant for our purposes.

Some research from the social sciences has already established classification schemes for gesture analysis [9, 4]. Even though these classifications usually use a slightly different definition of the word "gesture", we took them into consideration in our classification of gestures for manipulating graphic objects.

Most research on gestures, however, has focused on gestures as a substitute for spoken language or as a pre-verbal means of communication [17, 7, 12, 1], which is too far outside our interests to be useful in this experiment. This is true because our definition of gestures emphasizes object manipulation rather than interpersonal communication.

4 Method

4.1 Subjects

Forty-eight subjects were recruited from Carnegie Mellon University and University of Pittsburgh including students and staff for an experiment in "computer communication using speech and gestures."

4.2 Configuration of the Experimental Apparatus

The subjects were taken to the User Studies Laboratory at Carnegie Mellon University, where they were seated in front of a high-resolution, monochrome graphics terminal. Three video cameras and a microphone recorded the subjects' gestures and speech. One camera was facing the subject from above the graphics display terminal, the other camera was placed about two yards to the right of the subject. The third camera was positioned slightly behind the subject for a diagonal rear view of the hands. See Figure 1 for a sketch of this setup.

The first camera recorded the gestures of the subject from a top view. The second camera enable a different (side-)view of the gestures and hand motions. The third camera provided an alternate view which was necessary for detailed analysis, when some parts of the gesture were obscured from the other angles. In addition, by providing a view of the graphics display screen, this camera documented the visual context (on the screen) of the subject's communication. An experimenter controlled the display presented to the subject from a terminal and keyboard in an adjacent room, while monitoring both sound and video recording.

Figure 1: The physical setup of the experiment

4.3 Procedure

Twelve subjects were used for pilot experiments. The remaining **36 subjects** were divided into **three groups** of twelve according to their background. [18] The first group consisted of persons that are reasonably **experienced with computers**. The requirement was that they have been working with computers for more than 2 years and have written programs greater than a hundred lines of code.

The second group was comprised of **computer-naive** persons. We expected the subjects in this group to exhibit significantly different responses in their computer interaction when compared to the first group. These were people who had written no programs or programs less than 100 lines and only knew at most one computer language.

The final group consisted of people with a **spatial** reasoning background. This includes people who have worked in a machine shop or as designers or architects. The criterion for these was at least 2 years of work experience as a designer, machinist or architect as well as a completed machinist's apprenticeship or Bachelor's degree in design or architecture. The latter group was selected as a sample of people who use spatial thinking and object manipulation in a skilled work environment.

After a brief introduction to the experiment, the subjects were given these instructions:

> We are trying to find out if computers can learn to understand gestures and speech used together.
>
> Imagine that the computer is as intelligent as a human, and that it can see what you do with your hands and hear what you say. You will notice there are three cameras and a microphone in this room, which will all record your actions.

There are two objects, labeled 'A' and 'B', displayed on the graphics display screen in front of you. Object A (on top) is doing something. Object B (on the bottom) is not doing anything. Now try to make object B do what object A is doing, using only your voice and hand gestures. In order for the cameras to see your hand, make sure your gestures are made within the marked area.

PLEASE NOTE THAT THE COMPUTER DOES NOT KNOW WHAT OBJECT A IS DOING. So you cannot tell the computer to make object B "do the same thing"!

When your intentions have been understood, the computer will make object B do what you told it to do. Don't worry that you have never communicated like this to a computer. Just do whatever you feel would be natural to communicate what you want.

If you have any questions, please ask the experimenter. Begin whenever you see the objects on the screen.

Each subject repeated the basic cube manipulation task of the experiment for all operations and in all communication modes.

There were **three communication modes** which each subject used. The order of using the communication modes was randomized for each subject.

- The first communication mode allowed **only gestures** but no words. Here the subjects were told that the computer can see them, without being able to hear them. Any kind of gesture was acceptable for the purpose of the experiment in this mode. Subjects were asked to perform their gestures in reasonably close proximity (within 20 inches) of the screen, so the video cameras would be able to observe them.
- In addition to performing the basic task using only gestures as a means of communication, the subjects were asked to perform the task using **only voice** as a communication channel. In this mode they were told that the computer can only hear them, but not see them. Anything they said was construed as part of their attempt to carry out the instructions in manipulating the object on the screen.
- A third interaction mode utilized both **gestural and verbal** communication with the computer. Subjects were free to use whatever was natural for them to convey to the computer what it was they wanted to do.

In each communication mode the subject had to perform seven operations on a cube (see Figure 2). The order of communication modes and the presentation of the operations within each mode was randomized for each subject.

The seven different **operations** the subject was asked to perform upon an object visible on the graphics display. These can be classified into three basic groups: *rotation, translation and scaling.* The order in which subjects performed the operations within each mode of communication was randomized.

- In the three **rotation** operations, the object was rotating about each of the axis in 3-dimensions. The subjects attempted to make the target object perform the same rotation.
- In the two **transposition** operations, the object was moving either left or right from its origin to a position off the screen. The subject would try to transpose the object along the same axis (transposition here implies a sliding motion, changing the location of the object).
- In the two **scaling** operations, the object was growing or shrinking from its original size to a target size. The subject would try to rescale object B to the size of the target object A.

Before the actual experiment began, the subject was given three practice trials. In each practice trial, the subject had to manipulate the cube (translating, rotating and scaling, once each) with both voice and gestures allowed. These trials were randomly selected from the total set of possible trials used throughout the experimental session. These practice trials were not recorded. They were merely used to ensure the equipment was working properly and the subject had understood the instructions.

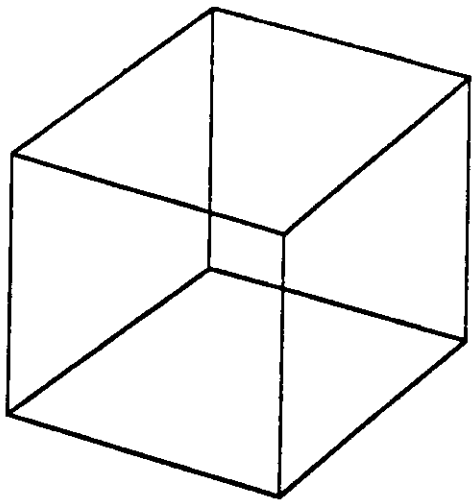

Figure 2: The actual cube used in all operations

The subject attempted to carry out all the different operations in all communication modes in randomized order. When the experimenter in the adjacent room judged that the subject had completed an attempt to set object B in motion, the experimenter caused object B to emulate object A via the control keyboard in the experimenter's room.

At the end of the experimental session the subject was asked to name a preference in the communication modes, he has just tried. In addition the subject was asked if he would prefer the typing to all of the modes in the experiment.

Further questions also allowed for general user feedback on the experimental setup as well as detailed criticisms or suggestions for each part of the experiment. Each session lasted about 25 minutes.

The structure of the experiment made use of a within/between-subjects design [16]. There are three groups (G) of twelve subjects (S) each, three communication modes (CommMode) and seven operations or trials (T) to be carried out, as shown below.

Other independent variables such as the speed of the object's motion size of the object and the distance of movement were held constant across all subtasks in this experiment. We also limited the type of object to a cube of fixed dimensions.

5 A Gesture Classification Scheme

We developed a classification inventory using the following features:

The main purpose of the experiment was not, to evaluate the *effectiveness* of the subject's gestures, but to determine empirically what various subjects will do in *attempting* to achieve the desired effect. Videotapes of the subjects were analyzed for several dependent variables. Two different raters were used to reliably analyze the videotape data according to a detailed gesture classification scheme developed in the course of this project.

	CommMode1							CommMode2							CommMode3						
	T1	T2	T3	T4	T5	T6	T7	T1	T2	T3	T4	T5	T6	T7	T1	T2	T3	T4	T5	T6	T7
	S1																				
G1	S2																				
	...																				
	S12																				
	S13																				
G2	...																				
	S24																				
	S25																				
G3	...																				
	S36																				

Table 1: The abstract experimental design

- The verbal communication of the subject during a subtask was classified and scored in several ways, as follows:
  - Number of words used in phrases describing the object's action or the intention of the subject
  - Number of words used in non-goal directed phrases. These include coughs, mumbling, speaking to oneself, and other speech events.
  - Number of lexical items, i.e the number of different words used. This can also be viewed as the *vocabulary* of the task.
  - The lexicon for goal-related words only. This lexicon has all the broken off words, or task-irrelevant words deleted from the vocabulary mentioned above.
  - Syntactic structure of the utterance, whether it represents imperative, declarative, interrogative or unclassifiable syntax.
  - Sufficiency of the speech act. This was a measure indicating if the utterance by itself could have conveyed enough information to allow the operation to be carried out. E.g. "Rotate" was considered insufficient, but "Rotate left" was considered satisfactory because it specifies the rotation of the cube to the left (around the y axis by default given the perspective of the user).
  - It was not possible to rate the real accuracy of the speech act. "Rotate left" could be viewed as perfectly accurate, given the right assumptions about default viewing angle or it could be completely inaccurate. We chose not to speculate about accuracy of the utterance for the intended effect at this point.
- The gestures of the subject were analyzed for the following features:
  - The identification and count of the fingers and hands involved at the start of the gesture. This measure describes the fingers and hands which were present when the gesture was initiated. In particular, it was recorded which fingers were present, whether they were extended straight or bent, and which hands were used.
  - Alignment of the gesture with the object on the screen. Frequently the gestures were not made directly on the screen in front of the object that was to be manipulated. From the videotapes the transcription recorded whether the gesture was shifted toward the subject, to the left of right of the object under manipulation and above or below the object. This alignment feature determined the difference in position between the "focus" of the gesture at the start and the object that was supposed to be manipulated. If one took a point from which the distance to all parts of the fingers that were used in

the gesture are minimized, that would define the center of gravity, or "focus" of the gesture. The gesture "focus" was an extremely useful aspect for identifying and classifying the gestures.

- The duration of the gesture. This was measured in tenths of seconds from the beginning of the gesture to its end. The beginning of the gesture was usually well defined, but the end was not always reliably identifiable. Thus this measure should be taken more as a ball-park figure for the gesture time.
- The number of repetitions of the gesture motion. Here the raters measured the number of times, the motion was repeated until the end of the gesture.
- The type of motion of the gesture. This motion was determined relative to the "focus" of the gesture at the start. We identified three basic groups of motions: rotation, sliding, and growing/shrinking. These corresponded fairly well to the types of operations of the objects: rotations, translation and scaling. A gesture was deemed "sufficient" if the type of motion of the gesture corresponded to the motion of the manipulation object on the screen.
- The identification and count of the fingers and hands in motion during the gesture. In contrast to the previous measure, a gesture may involve many fewer fingers in the motion than are present at the beginning. The fingers and hands that were in motion were identified here on the coding sheet.
- Looking at more detail in the gesture motion, the exact movement along each axis was also transcribed. In the case of rotation, the axis around which the fingers in motion seemed to rotate was identified. E.g. we said the rotation was around the x axis, clockwise. In growing/shrinking movements, the direction of the motion (i.e. inward = shrinking; outward = growing) was recorded. For a sliding motion, we recorded the direction of the motion and the axis (x = left, right; y = up, down and z = in, out).
- In addition to the fingers and hands at the start of the gesture, the palm position was identified, with respect to the subject. Thus the palm of each hand could face to the subject's left, right, up, down, toward or away from the subject. Finally, when two hands were used, the relative alignment of both hands with each other was recorded. E.g., the hands could be on top of each other or next to each other.

- Combinations of speech and gestures were further analyzed.
  - The time and duration of the gesture relative to the time and duration of the utterance was recorded. It was marked if the gesture occurred before the speech, during, after, before and during, during and after or throughout (i.e before, during and after) the speech.
  - Since the combined speech and gesture communication mode gave the subject a choice of communication channels, we also made a note which channels were used.

We have found that this detailed classification of gestures enabled us to reliably identify and code every gesture performed by the subjects. Conversely, we were also able to recreate the gesture fairly accurately from the mere textual transcription. The only aspect we found lacking in this classification was the magnitude of the motion. However, since we had held the magnitude of motion constant in all trials, our analysis of the data did not suffer due to this restriction. Therefore we feel we have developed an important tool for the analysis of gestures in man-machine communication.

6 Results

- Speech variables.
  - **Words.** Subjects used a total of 2397 words in 470 trials where speech was recorded. In the speech communication mode, they averaged 5.28 words per trial. Significant

	Avg Total	Rot. X	Rot. Y	Rot. Z	Expand	Shrink	Shift Left	Shift Right	Stat. Signf.
Avg. Words	5.28	6.36	6.16	5.13	3.88	3.77	5.97	5.63	<.001
Avg. Task Words	4.50	5.13	5.11	4.58	3.19	3.08	5.27	5.13	<.001
Lexicon	223 words	- - - - -			N.A.	- - - - - - - - - - - - - - - -			
Task Lexicon	141 words	- - - - -			N.A.	- - - - - - - - - - - - - - - -			
% Imperatives	93	94	94	97	94	94	86	88	<.05
Gest.-init. Fingers	3.8	2.7	2.8	2.8	6.0	5.5	3.1	3.1	<.001
Gest.-init. Hands	1.2	1.2	1.2	1.2	1.5	1.5	1.1	1.1	<.001
% Misaligned Gest.	78	75	75	91	80	83	68	75	<.05
Time/Gesture (secs)	1.7	2.1	2.1	2.1	1.6	1.6	1.4	1.4	<.001
Repetitions/Gesture	1.73	1.97	2.00	2.08	1.55	1.50	1.58	1.47	<.005
Fingers in Motion	2.19	2.47	2.44	2.05	3.83	3.83	.33	.33	<.001
Hands in Motion	.60	.11	.11	.27	1.13	1.02	.77	.80	<.001
Sufficiency	.94	.90	.90	.90	.98	.97	.96	.97	<.001
Sufficiency by comm. mode	gestures .96, speech .87, speech+gestures .98								<.01
Gest./Speech Sync. gest begin/end	during/during .5, during/after .29, before/before .02, after/after .04, before/after .06, before/during .09								
Stated Preference	.58 speech+gestures, .19 speech, .22 gestures								
Preference in Use	speech+gestures .70, speech .15, gestures .13								

Table 2: The Results at a Glance

differences were found between trials ($p < .001$)[2]. Translations required 5.97 and 5.63 words, scaling required only 3.88 and 3.77 words, while rotations consumed 6.36, 6.16 and 5.13 words.

- **Task-related Words**. Of these 2397 words, only 2007 were actually useful for the task. The rest represented inadvertent speech acts (hms, ahs), filler words (okay, yeah), mispronunciations and corrections. Again significant differences ($p < .001$) were found between trials. Scaling took the least 3.19 and 3.08 relevant words, translation required 5.13 and 5.27 relevant words, while rotations saw 5.11, 4.58 and 5.13 average task-related words per trial.
- **Lexicon**. The total lexicon used consisted of 223 words which included anything the subjects said, relevant or not.
- **Relevant Vocabulary**. The actual vocabulary for the useful words contained only 141 words.
- **Syntactic Structure**. Syntactically, almost all utterances were in the form of imperatives (93.6 percent). 2.9 percent of the utterances were declaratives while 3.4 percent were unclassifiable (e.g. "left"). There were significant differences in the percentage of imperatives used ($p < .05$). The transposition operations frequently used only the direction ("left" or "right") without a command verb. They only averaged 86 percent and 88 percent commands, while scaling operations averaged 94 percent and rotations averaged 94, 94 and 97 percent imperatives.

[2]For readers who are not familiar with the notation, "p" indicates the probability that these differences could have occurred by random chance alone

- **Speech Act Sufficiency** will be described below, with the effects of both speech and gestures.

- Initial Finger and Hand Positions:
  - **Fingers.** 1773 individual fingers showed up in the 464 gestures (3.8 fingers per gesture). Most often they were single fingers (29.0 percent), but also five finger gestures occurred frequently (24.1 percent) followed by two finger gestures 16.8 percent of the time and ten finger gestures (2.1 percent). Four finger gestures accounted for 8.6 percent of the gestures. There were a few six and eight finger gestures (.6 percent). The remaining 9.7 percent of the gestures only showed hands, but not individual fingers. Most of the time index fingers were present (97.1 percent), followed by thumbs (65.3 percent), middle (44.1 percent) ring fingers (40.9 percent). The little finger only showed up 40.5 percent of the time. There was a significant difference ($p < .001$) for the number of fingers used in each gesture trial. The three rotation trials showed only 2.75, 2.83 and 2.86 fingers on average, while the two scaling trials had 6.0 and 5.5 fingers average vs the two translation trials with 3.1 and 3.0 fingers average.

  - **Hands.** We saw 609 hands in the 464 gestures used in some way. The number of hands in the gesture mode trials also varied significantly ($p < .001$) for each trial. All three rotations averaged 1.2 hands, the two translations used 1.1 hands and the scaling trials averaged 1.5 hands per gesture.

  - **Alignment of Gesture with the Manipulation Object.** Of the 252 gestures in the gesture-only communication mode, 78 percent were somehow misaligned with the object on the screen. Again, this effect showed significant differences ($p < .05$) between trials. Translating left and right was misaligned 68.4 percent and 75.0 percent of the time. Expanding and shrinking operations were misaligned 80.5 and 83.3 percent of the time. The rotations were misaligned 75.0, 91.6 and 72.2 percent of the time. Depending on the axis of the rotation, alignment/misalignment was thus significantly different. Most of the misalignments, 69 percent were towards the subject, 42.4 percent were down or up, and 34.5 percent were left/right displaced. Many were misaligned in several ways (52.3 percent).

- Gesture time:
  - **Time per Gesture.** The gestures during a trial took an average of 1.7 seconds. There was a significant difference ($p < .001$) between the different operations, where the 3 rotation operations took the longest (2.1 seconds average), followed by the scaling operations (1.6 seconds average) and the translations (1.4 seconds).

  - **Repetitions of Gestures.** Each gesture in a trial was repeated 1.73 times on average. Repetitions varied significantly ($p < .005$) between trials, with the rotations averaging 1.97, 2.00 and 2.08 repetitions each, while translations had 1.58 and 1.47 vs scaling of 1.55 and 1.50 repetitions.

- Gesture Motions:
  - **Fingers in Motion.** There were an average of 2.19 fingers in motion during the gesture only communication. Again, differences between trials were significant ($p < .001$), with translations showing only .33 and .36 fingers per gesture on average. (Usually the hand would move shifting left or right, but no individual fingers). 3.83 fingers per gesture were average for both scaling operations and the rotation trials had averages of 2.47, 2.44 and 2.05 fingers in motion. Overall, 1070 fingers were encountered in a total of 464 motions, which includes both the gesture and the gesture and speech modes. Mostly single finger motions were encountered (27.1 percent), but also 5 finger motions (13.7 percent), two finger motions (12.9 percent), ten finger motions (8.4 percent) and a few three and four finger motions (together 6.6 percent of all gestures).

  - **Hands in motion.** An average of .60 hands in motion were found per gesture. Again

this differed significantly between trials ($p < .001$), from 1.13, 1.02 hands in the scaling gestures, .77 and .80 hands in the translations to .11, .11 and .27 hands in the three rotation trials. Note that only the hands that were moved during the gestures were counted for this measure. Thus in the rotations, many gestures were made involving only rotations of fingers, while the hand was relatively steady. In the translations, the motion of the finger was much more salient than the hand motion, which remained immobile with respect to the center of focus of the gesture.

- **Gesture Types**. We classified 199 gestures as rotations, 137 as translations and 128 as scaling gestures. Of all 135 gestures for translating objects, 132 (97.7 percent) were classified as translations gestures according to our scheme. The 125 gestures for scaling operations were matched by a gesture classified as scaling 98.4 percent of the time. The 204 rotation objects operated on with gestures were correspondingly classified with 96.0 percent rotation gestures. This result is also used in our analysis of "response sufficiency" below.

- **Gesture Motion Accuracy**. Unfortunately the data could not be analyzed adequately with respect to absolute perfect matches between gesture classification and desired object motion. This was due to the wire frame cube used in the experiment. Many subjects experienced "Necker-cube" effects but not quite) and were unable to decide on a uniform axis of rotation. They saw different axes of rotation depending on the way they viewed the cube. Looking at Figure 2 can illustrate how the cube could be seen in two different ways, according to which side the viewer decides is the front face. However, perfect agreement with the translation scaling operations and the gestures for them, was achieved 94.0 percent and 92.7 percent of the time, respectively. A match was scored when the desired object motion (growing, shrinking, shifting left, shifting right) was classified as a gesture of the appropriate type (scaling, translating) with the appropriate direction (inward, outward, left, right).

- Combining Gestures and Speech:
  - **Relative Timing of Gestures and Speech**. Gestures were made during the speech 50 percent of the time (i.e they began with or after the voice onset and ended before the utterance was finished). Gestures lasting from the beginning of the speech until after the utterance was completed were used 29.2 percent of the time. Gestures completed before speech onset occurred 1.6 percent of the time only, as did gestures done after the utterance ended (4.4 percent). Gestures throughout the speech (i.e. starting before speech onset and ending after the utterance was completed) happened 5.6 percent of the time. Finally gesture starting before speech onset and ending during the speech occurred 8.9 percent of the time. There were no significant differences between the groups or trials here.

  - **Preferences in the Questionnaire Answers.** In the post-experiment question, overwhelmingly ($p > .01$), 58.3 percent of the subjects preferred to use both speech and gestures. 19.4 percent preferred gestures alone and 22.2 percent liked speech alone best. There was no significant difference between the groups.

  - **Preferences in Use**. During the experiment, in 252 trials where a choice was possible, subjects used both speech and gestures 178 times, compared to gestures only 34 times and speech only 40 times. The only other significant differences were for trials using speech ($p < .01$) . Here subjects used speech most often in the scaling conditions (25 and 27 percent of all the scaling trials where a choice was possible). Speech was chosen for the three rotations 8, 8 and 16 percent of the time, while it was only used in translations operations 13 and 11 percent of the time.

  - **Sufficiency of the Subject's Response**. Even though perfect accuracy could not be measured for the gestures (see our comments on "Gesture Motion Accuracy" above), nor the speech (see "Speech Accuracy" in the classification discussion above), we did measure sufficiency of the response. In the case of speech, it was considered sufficient for the subject to say a verb and a direction/axis in any trial. In gestures, we

defined sufficiency as using the right kind of gesture (rotation, scaling or translation) for the respective operation. The individual axes or directions of the gesture were ignore. With these definitions, we found 94 percent of all trials were sufficiently performed by the subjects. There were significant differences between the seven trials ($p < .001$) and also between the three communication modes ($p < .01$). The gesture mode averaged 96.4 percent sufficient trials, while the speech mode averaged 87 percent sufficient and the combined communication channel reached 98 percent sufficient responses to the trials. The three rotation trials averaged 89.7, 89.8 and 89.7 percent sufficient, the translations were 96.2 and 97.2 percent sufficient to accomplish the manipulation and the scalings were 98.1 and 97.2 percent sufficiently performed.

7 Discussion

What have we learned from the speech data? We have seen a small number of total words, with few words spoken at a time. These results confirm the findings of Ford [5] who found that small vocabularies can be perfectly adequate for limited domain communication. Furthermore, these results also provide encouraging evidence that simple natural language grammars could be built to account for most of the linguistic communication. Finally, there are already commercially available speech systems, whose recognition ability exceeds the vocabulary necessary for this task [3]. Even though these systems are speaker-dependent and can only accept discrete words (i.e. the speaker must pause briefly between every word), that does not seem like a significant restriction given that the average number of words spoken was around three. So we have encouraging evidence that the speech aspects are indeed manageable in an integrated speech-gesture system of the future.

The large amount of data collected on finger/hand positions and motions gives rise to a few simple conclusions. People are not comfortable using only single fingers. Any system that restricts the user to a single hand or finger motion (like the "mouse") will be inadequate for many of the manipulations analyzed in this experiment. We feel it is safe to generalize to other manipulations in the real world, which are likely to be even more complex.

Some new questions were raised by the results showing subjects misaligning their gestures to the target object. In any gesture input system, the identification of the object under manipulation is of utmost importance. We can only speculate on the effects of feedback and on the ability of the users to modify their misalignment with feedback and practice. Presumably, immediate feedback will also affect the repetitions of a gesture. If the gesture was obviously correctly understood and the action carried out, further repetitions of the gesture will become unnecessary. In our experiments, the subjects only had limited feedback. There was no indication to them during the course of the gesture, where the computer thought their fingers were positioned relative to the object. Any useful gesture system will have to provide this feedback via the display.

The problem of judging the accuracy of a gesture, whether it conveyed the intentions of the subject, is also unsolved. Here too, the concept of feedback should provide a way for users to modify and adapt their gestures until the gesture recognition accuracy is adequate. This also ties in with other aspects of gesture recognition, not investigated by this experiment. The speed of a motion and the magnitude of the gesture can be easily corrected and adapted with adequate feedback and context information. Whether a sufficiently robust gesture recognition system can be built remains to be seen.

We also found that subjects made gestures in all three dimensions. They did not follow the outline of the screen but, depending on the effect they were trying to achieve, moved their fingers and hands around in all dimensions. Restricting them to a two dimensional input gesture would severely limit their ability to naturally communicate through gestures.

There is a lot of evidence that argues for the combination of input modes. Subjects preferred speech and gestures as an ideal combination in both use and stated preference. They also had better "sufficiency" in the specification of their intentions (in lieu of accuracy, which we were unable to measure) when the input modes were combined. Either mode by itself was less ideal. In particular speech alone made it difficult for

subjects to sufficiently specify all the parameters of the operation. Error checking and correction, perhaps with a form-filling program would be useful here. The results on the synchronisation of the speech and gestures make it seem plausible that in the presence of both, the gesture can be more reliably identified (especially the start of the gesture).

A big flaw in our experiment was the use of a wire frame cube. The resulting Necker cube effects, i.e. not knowing which face of the cube was in front, torpedoed some of our analyses. This was only discovered after the experiment was well underway. Unfortunately none of the pilot subjects had given us any indication of the extent of this problem. We do feel, however, that the experiment did reveal some basics principles for computer interaction with gestures.

There was a surprising amount of uniformity in the way subjects were communicating with both gestures and speech. This indicates that there are indeed intuitive, common principles in gesture communication. Our lack of significant differences between the groups of subjects confirms this. There are no expert users for gesture communication, it is a channel that is equally accessible to computer novices, experts and spatially oriented users.

One of the biggest achievements of these experiments was the development of a useful gesture classification scheme. This scheme is slated to become an important stepping stone in the analysis of gesture communication with a computer. We expect that many initial computer algorithms for gesture recognition will be based on the classifications provided in the scheme.

Many of the results presented above, are not meaningful in terms of statistical significance, but rather as a descriptive indication of the kinds of things people do when they communicate through gestures or speech and gestures. The interpretations of the results must be accompanied by a few notes of caution. The operations in this experiment represented only a small subset of the potential manipulations that could have been performed. Other operations not investigated in this project include constructing/assembling an object, exploring the magnitude of an operation (in the sense of a large vs a small-distance motion of the object, and the speed with which the action is performed fast vs slow-motion of the object). Obviously these data points could not give a complete picture of human-computer interaction using gestures. However, they provided first answers to many of the pending questions about manipulation of graphic objects using gestures and speech.

These operations represented only a small subset of the potential operations that could have been performed. Other operations not investigated in this project are constructing/assembling an object, exploring the magnitude of an operation (in the sense of a large vs a small-distance motion of the object, and the speed with which the action is performed fast vs slow-motion of the object).

Obviously these datapoints could not give a complete picture of human-computer interaction using gestures. However, they provided first answers to many of the pending questions about manipulation of graphic objects using gestures and speech. This experiment showed that gestures can effectively enhance computer input. The availability of an input device that is gesture-based can greatly increase human to computer communication bandwidth. Development of the gesticular language and the interpretive input device would be a great benefit in the application of computers.

8 References

1. Birdwhistell, R.L.. *Kinesics and context; essays on body motion communication.* University of Pennsylvania Press, Philadelphia, 1970.

2. Chapanis, A. Interactive Human Communication:Some Lessons learned from laboratory experiments. In Shackel, B., Ed., *Man-Computer Interaction: Human Factors Aspects of Computers and People,* Sijthoff and Noordhoff, Rockville, Md, 1981, pp. 65 - 114.

3. Dragon Systems, Inc. *VoiceScribe-1000: Speech Recognition System.* Newton, MA, 1986.

4. Ekman, P. and Friesen W.V. "Hand Movements". *Journal of Communication 22,* 4 (1972), 353 - 374.

5. Ford, W.R., Weeks, G.D. and Chapanis, A. "The effect of self-imposed brevity on the structure of dyadic communication". *Journal of Psychology 104* (1980), 87 - 103.

6. Friedman, L.A.. *On the other Hand.* Academic Press, New York, 1977.

7. Fromkin, V. and Rodman, R.. *An Introduction to Language.* Rhinehart and Winston, New York, 1978.

8. Furnas, G.W., Landauer, T.K., Gomez, L.M., and Dumais, S.T. "Statistical Semantics: Analysis of Potential Performance of Key Word Information systems". *Bell System Technical Journal 62,* 6 (July-August 1983), 1753 - 1806.

9. Greene, J.R.. *A Gesture Inventory for the Teaching of Spanish.* Chilton, Philadelphia, 1968.

10. Greene, R. "The drawing prism: A Versatile Graphic Input Device". *ACM SIGGRAPH '85 19,* 3 (1985), 103 - 110.

11. Hauptmann, A.G. and Green, B.F. "Comparing Command, Menu and Natural Language Computer Systems". *Behaviour and Information Technology 2,* 2 (1983), 163 - 178.

12. Hewes, G. "The current Status of the Gestural Theory of Language". *Annals of the New York Academy of Science 280* (1976).

13. Johnstone, E. The Rolky: A Poly-Touch Controller for Electronic Music. ICMC Proceedings, 1985, pp. 291 - 295.

14. Levinson, S.C.. *Pragmatics.* Cambridge University Press, Cambridge, 1983.

15. Martin, J.. *Design of Man-Computer Dialogues.* Prentice-Hall, Englewood Cliffs, 1973.

16. Myers, J.L.. *Fundamentals of Experimental Design.* Allyn and Bacon, Boston, 1972.

17. Paget, R.. *Human Speech.* Harcourt, Brace, Jovanovich, New York, 1930.

18. Potosnak, K.M. *Choice of Computer Interface Modes by Empirically Derived Categories of Users.* Ph.D. Th., Johns Hopkins University, Baltimore, Md., 1983.

19. Shneiderman, B.. *Software Psychology.* Winthrop, Cambridge, 1980.