Wizard of Oz for Gesture Prototyping

Jorik Jonker

Human Media Interaction Chair,
Department of Electrical Engineering, Mathematics and Computer Science,
University of Twente

Date:
July 10, 2008

Graduation Committee:
F. W. Fikkert, Msc.
dr. P.E. van der Vet
dr. D.K.J. Heylen

Student number:
0002291

Contents

Chapter 1

Introduction

Current computing has made a giant leap forwards in several areas over the past decades. Processing power, data storage, visualisation and connectivity have advanced almost beyond imagination. There is, however, one area where we still are stuck at the same level as in the beginning of personal computing: the input interfaces. In typical personal computing, a mouse and keyboard are both still an absolute requirement. For most applications, the mouse and keyboard do well enough to keep them in the picture, but for tasks like the manipulation of three dimensional objects, the traditional mouse has some shortcomings.

A default mouse has only two degrees of freedom[1] (*DOF*), whereas the human hand has six: three dimensional position and orientation, disregarding the fingers, which provide even more DOF. In order to employ mice in an environment where more than two degrees of freedom are needed, concessions have to be made, or the mouse is not suitable. Furthermore, when considering (very) large displays, using a mouse becomes ergonomically challenging (see Vogel and Balakrishnan, 2005). Finally, although using a mouse can almost be considered "natural" nowadays, one has to *learn* how to use it.

This study attempts to put some steps into shifting above paradigm by redesigning the interaction of a traditional application. An application in which spatial information is manipulated will be modified to be controlled by hand gestures only, since it is believed that this task could benefit from the shifted paradigm. This belief is supported by the fact that the two used map manipulation tasks have clear metaphors with physical manipulation. The digital analogy of a traditional map, the map application, was selected as program of choice. The map application shows a (large) map and offers two basic tasks: panning and zooming. Panning is the translation of the current view port to another location, while maintaining level of detail. Zooming is the act of changing the level of detail of the current view port on the map, without panning. It is believed that aforementioned tasks can be implemented with gestures using metaphors of a real, physical map.

1.1 Research question

The problem description mentioned above is summarised in the following research question:

[1] The scroll wheel can be regarded as a separate input device

> Which hand gestures make up an *intuitive* interface for controlling a map application?

The goal of this research is to prototype a gesture interface providing an intuitive interface to the map interface.

The rest of this report is organised as follows. The next chapter describes the overall methodology of this study, followed by parts I and II, which respectively deal with the design and execution of the experiment. Finally, chapter 11 will generally discuss the results of this study, answering the research question.

Chapter 2

Methodology

As mentioned in the previous chapter, the search for intuitive hand gestures will be supported by the map application. These gestures need to be "fed" into the computer, requiring both a tracker, capturing the gestures, and a recogniser. It is trivial that the recogniser needs a gesture repertoire, since it needs to know what to recognise. Figure 2.1 gives a schematic overview of this gesture controlled application.

To avoid having to implement these tracking and recognition techniques, which are both quite complex, a human operator will be used to fill in those tasks, while prototyping. This method of (secretly) employing "human computing" is called a *Wizard of Oz* setup (see Kelley, 1983a,b), where the end user does not know that some parts of the system are actually performed by an operator, or *wizard*.

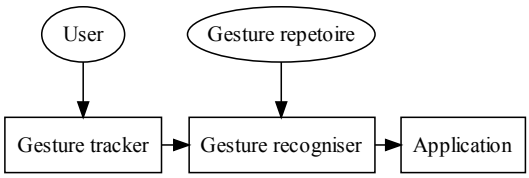


Figure 2.1: Block diagram of the application

The general methodology of this research was identified as follows:

- The gestures will be "extracted" from the subjects by simply letting them interact with the application;
- They will be given several assignments which require the map application to complete;
- This session will be registered in a way allowing the processing and extraction of the gestures afterwards;
- The gestures will be annotated in such a way the research question can be answered.

In order to commence such a session, the methodology of the session itself should be very clear. The design of this methodology is dealt with in part I, which describes an experiment with the map application. The practical setup of the session as well as an annotation scheme will be covered in this part.

Part II will deal with the execution of this session. Moreover, the observed gestures will be discussed in this part.

Part I

Experiment design

Chapter 3

Introduction

This part of the research deals with the design of the gesture session. The session consists of an interaction between a test subject and the map application. The goal is to create an intuitive interface, which in this case would be satisfied by creating an interface which requires no, or minimal adaptation (i.e., learning) from the end user. In an attempt to realise this intuitivity, the interaction will be defined by the observed gestural behaviour. The user will not be instructed on how to interact with the application, except that it should be done with hand gestures only. Sessions with users will be recorded on video, such that later analysis can provide an interaction programme for later revisions of the application.

Since the implementation of a gesture recognising and tracking system is beyond the scope of this research, this part is simply replaced by a human operating the application (the *wizard*). There still remains one problem, however, namely the "programming" of this wizard: how the wizard should react on his observations.

Since there is no information available on the semantics of the observed gestures, the presence of the wizard will be disclosed to the end user. Furthermore, the user will be encouraged to speak out loud his intentions, so that the wizard is able to operate the application accordingly. This verbal "side channel", which can be used to correct misinterpretations of the wizard, is believed to overcome the bootstrap problem of the wizards programming.

3.1 Research question

This section describes the skeleton of this part of the research, formalised in a research question, which will be decomposed into several sub questions. These sub questions are grouped into two main pillars: suitability and analysis. The main research question of this research was formalised as:

Is the Wizard of Oz paradigm suitable for obtaining gestural repertoires?

In the first place, this study addresses the suitability of the human factor in gesture interfacing. The acquiring of gestural repertoire will be dealt with during the analysis phase. If all of the questions below can be answered positively, this session is followed by an intrinsic study dealing with the analysis of the actual gestures and their semantics (see part II).

1. Does the presence of a human operator introduce no significant extra latency?
2. Is the operator able to track the subject correctly?
3. Does the subject feel “in control”?

Furthermore, this study will determine an analysis method for the video material. In a larger perspective, the video material needs to be compared with each other, which can be eased by a form of abstraction. An annotation suits this purpose, so the video’s will be annotated. We do not want to annotate unnecessary details, since this will have negative impact on the analysis. A textual representation of the video material would serve this purpose quite well, so an annotation scheme is to be searched. It is well known that the annotation of video material is a very labour-intensive task, some kind of optimisation would be very welcome. If subjects show internal consistent gestural behaviour, for example, the video’s could be grouped and only partially annotated, which would save considerable time. If this study shows that the gestural behaviour is individually consistent, the next part could benefit from this optimisation. This leads to the following questions:

4. How should the video material of the experiment be annotated?
5. Do the individual test subjects show individual consistent gestural behaviour?

Chapter 4

Methodology

This chapter deals with the exact methodology of this part, describing the session and its analysis in detail.

4.1 Session setup

The first thing to be specified is the overall setup of the session, which consists of several parts: the location, the used tooling and the procedures. The next sections describe each aspect of the session.

4.1.1 Location

The session requires a large screen, which was implemented by using a digital projector, connected to the PC running the application. The motion capture lab of the University of Twente was chosen, because of the availability of a sufficiently large projection screen, a permanently mounted projector and because it is a relative large room, which allows a lot of working space. The permanency of this projector is valuable, since it encourages more consistency across the sessions. The room is partitioned in two areas: an elevated (square) floor in front of the screen, and a "control bridge", consisting of fast workstations connected to the projector. The elevated floor measures roughly 60 m^2, while the projected screen is 25 m^2. From the control bridge, one has a clear view on the elevated floor and thus the end user, during the session. The whole lab can be darkened using curtains at will, in favour of a good projection quality. Figure 4.1 shows a schematic overview of the setup of these sessions.

The elevated floor appears to feature a marked square in the middle, which will be used to ensure that the user's position in the room across sessions was consistent. There is no need to "hide" the operator from the end user (which is usual in a classic Wizard of Oz session), the end user was even encouraged to have verbal contact with the operator during the session.

Finally, the sessions will be recorded on video. It is chosen to use only one camera during the session, since multiple cameras could significantly raise the analysis' complexity. The position of this one camera, however, has to be determined using simple trial and error. Figure 4.1 shows the different options for this camera position: one from the rear, which approaches the view of the operator; one from the front, having a clear view on the user's hands and one from the left corner, having optimal lighting

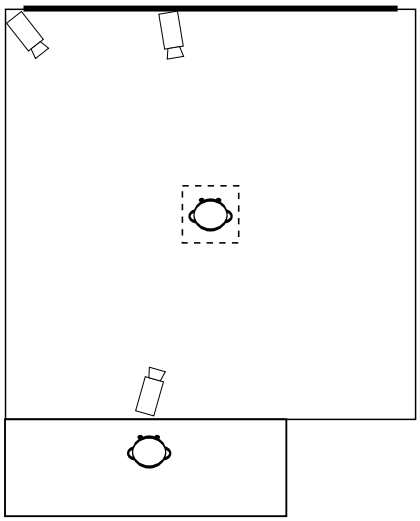


Figure 4.1: Top view of the setup of the session. The dotted square in the middle of the bigger square is the marked area in the middle of the elevated floor with the end user standing in it, facing the screen (thick black line). The three camera-shaped objects are the possible camera positions, the rectangle on the bottom is the control bridge with the wizard in it.

conditions, since the left has the main light source of the room. To make the decision which camera position is the most suitable, the videos are viewed and compared afterwards, keeping in mind the analysis task. This decision could influence the way the operator has a view on the subject, which was evaluated after the session.

4.1.2 Application

The application will be developed as an "extended image viewer", since most image viewers cater the two tasks specified in the previous chapter: panning and zooming. There are, however, some specifics, which make the search for a suitable image viewer unfeasible in favour of developing a small (Java) application which does meet the requirements:

- full screen image viewing;
- hidden mouse;
- being able to log or record the current view port;
- being able to reset the view port to preset positions;
- having very fine grained control on the interface for the operator.

The first two of these requirements enhance the immersiveness of the application, since when fulfilled, the only visible thing on the screen is the map. The next item can provide crucial information during the analysis of this session, since the recorded video can be paired with the screen contents, afterwards. The ability to reset the view port to preset positions makes it possible to start an assignment in a fixed position. The

last requirement is a very important one, since the ease with which the operator is able to control the application has direct impact of the latency introduced by the human operator. It is trivial that the lower the latency introduced by the operator, the more immersive the experience when using the application.

The application is developed using out of the box methods for image viewing and scaling. This custom development will allow fine control of the operator interface: the application will feature a "grab and drag" mouse motion to pan the current view port, while the scroll wheel of the mouse will be used to zoom the map. The visual (projected) output of the application will simply be the current view port on the bitmap, stretched to fill the whole screen. The bitmap loaded into the application will be a high-resolution topographical map[1] of Twente. The session will provide feedback for the enhancement of the map application in the next iteration, since it would be the first time the application was employed in practise.

4.1.3 Session Intrinsics

The session itself consists of an introductory talk between the host (mostly the researcher or operator) and the end user, several phases in which the user was given different assignments, and finally an evaluation. In the introductory talk, the host explains the functionality of the map application: its two features (panning and zooming) and the fact that it needs to be controlled by hand gestures. Furthermore, the host does not explain *how* to interact, but rather encourages the user to just try to interact with it, speaking out loud the intended actions. Finally, the user will be told that if he did not understand something, or had trouble solving an assignment, he could consult the host.

The first assignment phase of the session encompasses getting acquainted with the application and the operator. The user will be asked to simply play around with the application to feel how it works, while the operator becomes familiar with the gestures of that user. This phase is ended by the operator, giving the user his first assignment (typically after a minute). The first series of assignments consist of positioning the view port in such a way that a given city is centred and filling more or less the whole screen. When the user indicates that he does not know that location of that given city, the host could provide hints (for instance: "Delden is located to the North West of Hengelo"). Each subject has to complete the same three of these assignments, which are provided at random order. This basic assignment is chosen, because it does not require very precise control.

The second series of assignments is like the first series, but this time the view port needs to be positioned around three given cities, in such a way that they all three are on exactly the edge of the view port. The idea behind this assignment is that it requires a somewhat finer grained control of the application, and thus involves more smaller zoom and translate movements.

The last assignment is not a series, but a single one: the user will be asked to find a random place of personal interest. It is believed that these series can be biased by the fact that the operator was knowledge about "solution" of the assignment. Since the operator does not know which view was chosen by the end user, this bias could not occur.

In order to answer the first three research questions, which evaluate the end user's experience, a small evaluation will be done at the end of the session. The user and operator are simply asked to answer the three research questions.

[1] The map is what the Dutch refer to as "stafkaart"

The session will be done with two people involved with this research. It is expected that these people were biased by their involvement to this experiment. Sessions are grouped in trials, in which the role of operator and user will be interchanged. In total, there will be a number of three trials: one for every camera position.

4.2 Analysis

The next point of focus will be the analysis of the session. In the first place, an annotation scheme needs to be developed. This annotation should translate the video images into some gesture notation. In order to determine a suitable annotation scheme, the next paragraph discusses the requirements of that annotation formalism.

The main requirement is quite trivial: being able to transcribe the gesture utterances in the videos. It is believed that two gestural utterances from different persons can be both considered different and the same, depending on the level of detail in which they are regarded. The level of detail determines the differences and similarities between two gestures. The formalism was thus required to provide a means in which the level of annotation detail could be defined.

Since the main part of interest is the gestures and their corresponding actions in the application, only the parts of the video containing "zoom" and "pan" acts are of interest, which allows to skim the videos somewhat. Figure 4.2 shows a decomposition of the video material. On the first level, the different sessions are segmented (in practise: in separate media files), where the next level segments the different tasks. Movement epenthesis (Ong and Ranganath, 2005) (*M.E.*) are the inter sign transition periods, what in this case means that gestural utterance is not of interest. The third level makes a distinction between the two implemented tasks (panning, zooming and again M.E., which is not considered), while the fourth distinguishes between different gesture phases (preparation, stroke and retraction). The actual annotation will take place when transiting from the fourth to the last level, where the strokes are translated into a written medium. A proper gesture transcription language is needed for this last step.

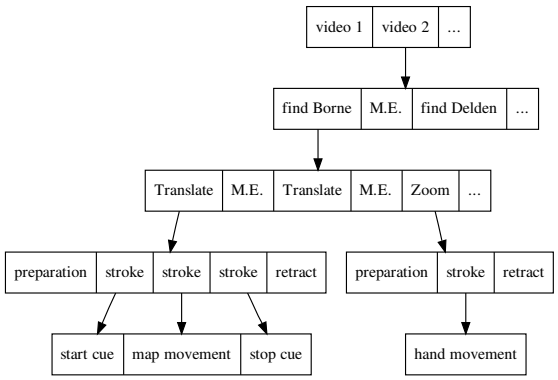


Figure 4.2: Video decomposition

The field of gestural research features three important gesture transcription languages: Stokoe (Stokoe et al., 1965), HamNoSys (Prillwitz et al., 1989) and SignWriting (Sutton, 2007). All three formalisms are developed to transcribe sign language, with Stokoe focusing on American sign language (ASL) specifically. HamNoSys is

the most expressive of all three, which is also reflected in the complexity of the language. While HamNoSys and SignWriting are pictographic formalisms, Stokoe codes utterances into Latin characters. SignWriting is, compared on complexity and ease of use, positioned between the other two.

In order to decide between the three languages, trial and error will be employed in attempting to annotate the strokes in the three languages. Stokoe will be attempted first, since it has the greatest ease of use, due to its "normal" vocabulary. When Stokoe does not suffice, SignWriting will be used, since it is the least complex of the two pictographic languages. Finally, when SignWriting also does not work, HamNoSys, being by far the most challenging language to use, will be deployed. The first language able to annotate the videos "wins".

Finally, the annotation could benefit from the knowledge that users show internal gestural consistency: if, for instance, a person uses consequently the same gesture for a certain task, those utterances can be grouped. Only one annotation per group is required in that case, which would save considerable time. This hypothesis can be tested by grouping gestures that appear similar (without looking at the annotations) by hand, looking at their similarities and annotating each group member afterwards.

In order to facilitate the annotation of the session, some form of tooling is needed, meeting certain requirements. In the first place, the tool has to support multiple annotation tracks, so the video can be annotated on several levels. Furthermore, it has to be able to use current codecs like *MPEG-4*, allowing the video's to be compressed to a reasonable size. Next, a multi-platform tool was preferred, since this allows annotation in a heterogeneous environment. A quick search on the Internet learns that Anvil (Kipp, 2001, 2004), an annotation tool written in Java, meets these requirements.

Chapter 5

Results

This chapter describes the results and observations of the session, commenting on the practical aspects of the session, the experience of both user and operator and finally the analysis. The interpretations, discussion and conclusions of these results are posed in the next chapter.

5.1 Practical aspects

This section deals with practical variables of the session, which needed to be determined. At first, the effects of the different camera positions are discussed, followed by feedback on the application interface.

5.1.1 Lighting conditions

In order to have a decent view on the screen for both operator and subject, the light in the lab had to be severely dimmed. On the other hand, in order to get usable video images, the lighting had to be as bright as possible. Since these two preferences are in conflict with each other, the matter was settled by almost closing the curtains of the lab, leaving a gap of 10cm across the width of the room. This allowed a decent view of the screen, and a video quality, which was just good enough for processing after some enhancing steps. These steps consist of boosting the brightness and contrast of the video's, since in the raw video's it was hardly possible to distinguish the subject from the background. After processing, the videos still were of mediocre quality, but the gestures were distinguishable.

5.1.2 Camera position

The same sessions were done using three camera positions, to decide which was the most optimal with respect to analysis afterwards. Having seen all six video's, it appeared that the images from the camera right in front of the user contained the most detail, since that position showed the most visual information about the subject's hands, which eased the analysis of the session. The position from the rear had a view approaching the operator's view the most, but the user appears on this video merely as a dark shadow, in front of the (bright) screen. Finally, the position from the front cor-

ner indeed had relatively good lighting conditions, but had a bad view on the subjects hands.

One of the two operators had indicated that the session could benefit from a view from the same position as the camera, which reveals more of the user's hands.

5.1.3 Application

The first thing that came forward about the application was that there was no visual mean to provide the user with feedback when the view port reached the "border of the map". When the view port was translated to the border of the bitmap containing the map and the user attempted to move beyond the border, the view port seemingly "froze", displaying the last displayable view.

Furthermore, the operators indicated that the "grab and drag" paradigm could probably be replaced by a more suitable one. It was suggested several times to implement a means with which free mouse movement was directly mapped to map translation. This mapping could be triggered by a key press, so that the mouse was not always "coupled" to the map.

Finally, the application suffered from a minor bug, triggered by the fact that the workstations in the lab had multiple monitors. Since the mouse cursor in the application was hidden, the operator could not see if the mouse left the current screen. If that would happen, followed by a mouse click (because the operator tried to drag the view port), the application lost mouse focus and minimised. It is believed that this bug did not severely impact the session.

5.2 Experience

After the session, both user and operator were briefly and informally interviewed on their experience with the application. This interview dealt with the latency, the ease of tracking by the operator and the extend toward which the user felt "in control". In all of the six sessions, the users never have indicated to notice or get annoyed by the amount of latency, introduced by the human factor. Apart from those aspects, both subjects were positively surprised by the immersiveness of the application. The expectation of both subjects was that the presence of the "wizard" in the interaction would be far more obvious than they experienced.

During the first sessions, it was obvious that the operator needed to become acquainted with the tracking of the user, which became more and more routine during the later sessions. Both operators incidentally "mirrored" the user's actions: when the user tried to move the map from left to right, it moved in the opposite directions. This mirroring occurred with both acts: panning and zooming. Both subjects indicated to be surprised by the immersiveness and level of feeling in control.

5.3 Annotation

After being enhanced, the video's were annotated on the second level: assignment. The first level of segmentation was already done by the person capturing the video's on hard disk. This yielded seven elements per video: one sync and two sets of three "real" assignments. These elements were drawn from a set of three possibilities: SYNC, "Find one city" and "Find multiple cities". It was decided to have a separate file per

session, so the first track within an Anvil file are the assignments. This annotated level did not cover the whole video, since there were no parts of interest between the assignments (M.E.).

On a second track, the real assignments were split into the *tasks*: pan and zoom. These two tasks were annotated only for the real assignments e.g., not during SYNC or M.E.

The third track distinguished the gesture phases: *prepare*, *stroke* and *retract*. Again, these were only annotated in the parts were the previous level served an annotation, since we are not interested in the details of the motion epenthesis or SYNC.

The fourth track, dubbed *transcription*, contained free text elements in which the gestural details of the previous two tracks could be transcribed. As mentioned in chapter 4, Stokoe, being the least complex of the possibilities, was used in attempting to transcribe the recorded gestures.

5.4 Gestures

Before going into the details of the transcription of the gestures, this section will give an impression of the observed gestures. It appeared that the gestural implementation of the tasks are very similar across the two subjects.

The pan gesture was implemented by a strike of the arm, moving in the vertical plane parallel to the screen. This displacement is illustrated on the accompanying CD by the file named `typical pan.avi`. This strike is a mapping of the intended displacement of the map, as if the (virtual) map was grabbed, dragged and released. To indicate the start of the drag, the subjects had their own cues, which was the case for the end cue as well. Subject 1 signalled the start of the pan by spreading the fingers, keeping the hand spread during the whole task. At the end of the task, the hand returned to a normal "flat hand" state. Subject 2 signalled the start of a pan by closing the hand to such a position that only the index and middle finger are stretched. Occasionally, the middle finger was left closed as well. During the pan, the hand kept this state. The end of a pan task was signalled by the hand opening up to a "flat hand".

The observed zoom gesture was a two hand gesture, in which both hands performed the same movement symmetrically. The gesture was signed in front of the body, with both hands pointing from the signer. When zooming in, the hands started close to each other, with increasing distance between them, indicating the increase of detail level. When zooming out, the inverse of this gesture happened: the distance between the hands decreases. These zooms, respectively in and out, are exemplified on the CD by the files named `typical zoom in.avi` and `typical zoom out.avi`.

5.4.1 Stokoe

In order to use Stokoe together with Anvil, without modifications, the ASCII variant of Stokoe, ASCII-Stokoe (Mandel, 1993) was used. Aside from its convenience by using only "typable" characters, it has some extensions over regular Stokoe, like a few more hand shapes and a more consistent notation.

One of the first observations was that Stokoe allowed a transcription of most of the observations. In order to make the gestures "fit" into the language, some details had to be discarded. An example of this abstraction is the amount of available hand shapes, of which 19 are covered by ASCII-Stokoe. Another example is the fact that Stokoe only covers a limited subset of motions and orientations, although these can be combined

infinitely. Despite this lack of detail it was believed that ASCII-Stokoe was powerful enough to annotate the gestures in sufficient detail. This issue will be discussed in chapter 6.

An interesting property of Stokoe is that the language features no distinction between the left and right hand. Motion and orientation towards the left or right are implemented as towards the dominant or non-dominant side. The dominant hand is the signing hand, which leads to ambiguities when two hands are used. It appeared to be, however, that in all the zooms and pans of the sessions of both subjects either one hand was used, or two hands in a symmetrical way.

It appeared that Stokoe mainly facilitates locations on the body itself, since American Sign Language is mainly signed on the face, on another hand, on the body, etc. Stokoe does not cover the "free air" locations observed in the video's. This was annotated by using the *neutral* location, `Q` for every sign, which mains that the sign was performed in front of the signer. According to Stokoe tradition (see Stokoe et al., 1965), this `Q` is often used when it does not matter very much where the sign is signed.

It appeared that the observations could all be split into three blocks, not completely unlike the well known *prepare*, *stroke* and *retract*. A typical pan gesture of both subjects involves the subject changing its hand shape into an active shape, followed by a free movement in the air, finished by change in hand shape into a neutral shape. Each *zoom* gesture followed this same paradigm as well. This was implemented in Stokoe using three separate transcriptions. These three gestures were interpreted as a simple state model (see figure 5.1), with two states: track and neutral. In the track state, the user intends the system to track the motion of the hand(s), e.g., in a pan task the map would be displaced linearly to the user's movement. The first gesture of these three is considered as a cue to start tracking, whereas the last represents the cue to stop tracking.

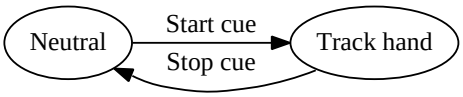


Figure 5.1: Gestural states for typical pan and zoom tasks

The free movement in the second sub-gesture was often quite complex, while Stokoe covers only (very) basic movements, although these can be combined infinitely. Semantically, these movements all had a consistent meaning within their task: within the pan tasks, a free movement meant that the map was to be displaced linear to this movement. The same semantics applied to the free movements involving the zoom acts. The choice was made to not describe these specific movements in detail, since it was believed that this tracking state did not contain any information besides (hard to transcribe) complex movements. This resulted in a minor extension to ASCII Stokoe: the addition of two "free" movement symbols[1]. Since both subjects showed very similar gestural implementations for pan and zoom, they could share these movement symbols. The pan task was typically implemented by a start cue, followed by a movement in the plane parallel to the screen (up/down, left/right), followed by the stop cue. This was annotated using the new invented `&` sign, which means that there is a movement of the hand in that plane.

Furthermore, the zoom task was implemented by both subjects a symmetrical movement of the hands, either towards or from each other, for zooming and out, respectively.

[1]These symbols were drawn from the set of unused symbols, as specified in Mandel (1993)

The motion of this zoom task was annotated by the new `Z` sign, which means that there is a movement of the hand, either towards or from the other hand, parallel to the horizontal axis. The circumfix `S(...)` indicates that a gesture is performed by both hands in a symmetrical way, like both subjects did when doing the zoom task.

To ensure a consistent annotation across multiple sessions, the interpretation of Stokoe and its specific extensions are documented in a annotation manual, see appendix A.

Table 5.1 shows some basic statistics of the Stokoe annotations.

Table 5.1: Gesture counts

(a) Unique gesture counts

subject	pan	zoom
1	16	8
2	61	8

(b) Total gesture counts

subject	pan	zoom
1	81	30
2	88	13

Chapter 6

Discussion

This chapter discusses and interprets the results provided in the previous chapter.

6.1 Practical Aspects

It was decided that the camera position in front of the screen was to be used in later sessions. The camera, however, receives a different image than the operator, who is situated at the back of the experiment. The camera, for in stance, had a far more better view on the user's fingers and hand motion than the operator. The result of this could be that the analysis and the operator do not get the same information, which could bias the analysis.

The feedback on the application has lead to a list of enhancement for a further iteration of the application. The first of these is a keyboard command, indicating that mouse motion has direct control on panning and zooming. This could be implemented by using a modifier key, when pressed dictates that mouse movement is mapped to map movement (or zooming).

The next improvement is the implementation of visual feedback on the map's borders. This could be done by adding a distinctively coloured area around the map, indicating that the map has ended. When this is done to the bitmap containing the map, it did not require making alterations to the program, keeping it more simple.

The bug, triggered by a dual screen setup could be hard to tackle in the application itself, since the core of the problem lies somewhere "deep" in very platform-specific parts of Java. Instead, it could be tried to disable the second screen in the lab, working around the program. This was explicitly not done during the trial, since this would be a change to the (very complex) lab setup, which is used by other research groups as well.

6.2 Experience

The results of the session were very positive, suggesting that the form of the interaction (gestures, using a wizard) are suitable for gesture search. A point of focus for next sessions could be that it is very easy for an operator to accidentally mirror the user's actions. Both operators could not come up with a specific reason for the mirroring.

It appeared that the presence of the wizard did not induce unexpected behaviour. When the presence of the wizard in the interaction process would not be disclosed to

the user in further sessions, this could induct inconsistent gestural behaviour between the research parts.

6.3 Gesture Stages

The previous chapter has introduced the term "stages" for the transcription of pan and zoom tasks during the first session. The first stage of the gesture signals the start of the second, while the third stage signals the end of the second stage. One could state that this second stage has the closest link with the semantics of the gesture, not unlike the stroke of a gesture. According to McNeill (1992), the stroke of a gesture caries the *imagistic content* of a gesture. Since the first and the third stage is purposely and very observably uttered, they are regarded as being part of the imagistic content of the gesture and therefor not as a gesture phase. Besides, they are both "surrounded" by a real prepare and retract phase, which do not appear to have this imagistic content.

6.4 Stokoe

Although Stokoe was used to successfully transcribe the first session, it has some drawbacks. The biggest point of discussion is the level of detail, covered in Stokoe. In the first place, Stokoe only covers 19 distinctive hand shapes. Any other hand shape can simply not be expressed in that language. While this may seem a little bit restrictive, one has to keep in mind that "reality" covers infinite hand shapes, so some form of abstraction has to be used when using any language, in finite space and time. Using Stokoe, some gestures needed to be fit into Stokoe's hand shapes, loosing some details.

Another detail getting lost in the translation to Stokoe is which hand is used for a gesture (left or right). Stokoe uses the convention of dominant or non dominant hand, which abstracts upon the hand used. Moreover, it could lead to ambiguities, when transcribing two-handed gestures. There is, however, a possibility to enhance Stokoe in such a way that this detail is not lost. In the previous chapter, the circumfix `S(...)` is introduced to denote a "dual hand" gesture, which can be extrapolated to `L(...)` and `R(...)`, to denote a left or right handed gesture.

Regarding the main research question, which focuses on the *typical* gestures, the fact whether a gesture was uttered with left or right is not very interesting, if this handyness has no semantic meaning. Since the semantics of this handyness are currently not researched and thus unknown, it is decided not to follow up on this matter. The videos support this decision, since they do not suggest semantics coupled to the handyness.

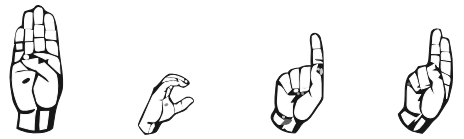
Furthermore, the introduction `&` and `Z` motion symbols is responsible for a big abstraction. These two symbols reduce (optionally complex) motion into a simple symbol. While this abstraction allows for a very efficient annotation, it leaves out potentially valuable information at a very early stage. However, in a broader context of this research, the exact motion during pan and zoom tasks does not contribute at all. An alternative to this approach is to describe these motions in full detail (probably involving the magnitudes more complex HamNoSys language), after which an abstraction is done, in which the motion could be discarded. The latter approach would be much more elegant, but also extremely time consuming.

6.5 Abstraction

The previous chapter shows that although both subjects visually show quite similar and consistent gesture behaviour, their annotations are quite diverse. Since the goal of this study is to provide a set of *typical* gestures for a set of given tasks, a reduction in annotation detail would be very welcome. The gesture terminology in the next sections is explained in appendix A.

To decrease the number of different annotations per task per person, an *ad hoc* abstraction needs to be developed. Table 5.1 shows a lot more than 2 unique gestures per subject; one for every task, as suggested in section 5.4. A lot of non-uniqueness can be reduced by increasing annotation consistency: using a fixed order (alphabetical) and leaving away implied modifiers. Moreover, if motion implies an ambiguous change in hand shape, an "end" hand shape should be dictated.

The fixed order of modifiers ensures that `Q/B/v,>` and `Q/B/>,v` can easily be recognised as the same annotation. (ASCII) Stokoe does not dictate a specific order, so the ASCII order of the used characters will be used when there are multiple possibilities. This was done using a simple computer program. Leaving away implied modifiers means that when motion implies certain modifiers, these implied modifiers should not explicitly by annotated. This may sound very trivial, but it appeared that a lot of these implied modifiers are still present in the annotations. For example, a `&` (pan) motion starting with the `B`-hand pointing to the dominant side, making half a circle ending pointing to the non-dominant side implies pronation (`b`), which can be seen in file `pan pronate.avi`. Moreover, the `&` motion can imply the motion characters >, <, ^ and `v`, since it described motion in the plane parallel to the screen.



(a) "B" (b) "C" (c) "G" (d) "H"

Figure 6.1: Hand shapes

Furthermore, it also appeared that there was confusion about the correct motion representing the transition between two different hand shapes. On a rare occasion, the transition between the `B` to the `C` hand (see figure 6.1) was marked by the symbol `]` (open), whereas it is obvious that this should be a `#` (close). More frequently, the transition from a `C` → `H` and `C` → `G` was marked alternating by `#` and `]`, which are less trivial to disambiguate. Since the hand partly closes *and* opens during these motions, their occurrence counts were simply consulted in order to determine which is "correct". For both `C` → `H` and `C` → `G` it was determined that `]` is the right motion symbol

Finally, the end hand shape in the motion part of Stokoe should only be used when the motion implies an ambiguous change in hand shape. The motion `#` (close) and `]` (open) are good examples of this, since the motion does not exactly specify *how much* the hand is opened or closed.

The real reduction of unique gestures was obtained using some very simple generalisations, where less interesting details (i.e. without intended semantics) were left away. The basic idea of these abstractions was that every annotation detail without intended semantics should not be reflected in the annotations. For example, all observed pan gestures reflect the metaphor of a physical map being dragged by the motion of

the user. The user intends to "grab" the map, which is commonly represented by #, the symbol for closing the hand. In some occasions, the user also pronates or supplinates the lower arm, represented by the symbols a and b, but the videos suggest that this motion is merely a side-effect of the main motion. Going on with this example of the pan metaphor, when looking at a close enough detail, almost each pan is preceded by a directed motion in a direction unrelated to the direction of the intended displacement, as if the user positions the hand to some "starting position" before the actual pan commences. This fact appears to be a main cause of a multitude of different annotations of gestures visually appearing the same.

In order to reduce the details not representing intended semantics, this directional movement in the starting gesture phase is simply removed. This movement, however, is often reflected in a end handshape, so it has to be cascaded in order to maintain consistency. After this canonalicalisation and these two abstractions, the unique gesture counts were reduced to the figures displayed in table 6.1(a).

Table 6.1: Gesture counts after abstractions

(a)

subject	pan	zoom
HMIG1001	2	4
HMIG1002	23	6

(b)

subject	pan	zoom
HMIG1001	2	2
HMIG1002	4	3

The next item addressed was the abstraction of the starting handshape of the first gesture stage and the ending handshape of the third gesture stage. It appeared that although the observations of subject HMIG1002's pan gestures look very similar to the human eye, the annotations were still quite diverse. The major issue seemed to be the very first and very last handshape of the gestures.

Most gestures are annotated as starting with a "relaxed base hand", which was annotated as a C. There is however, a small group of gestures which is annotated using a different hand than this C. A closer look at the videos learnt that the observed hand shapes are technically between Stokoe's B and C, leaving it to the annotators preference to decide between these. Since by far the most gestures start with a C and end with it too C (117 of the 169 observations), it was decided that each gesture starts and ends with a C.

The remaining unique gesture counts of subject HMIG1001 are displayed in table 6.1(b). Subject HMIG1001 shows very consistent gesturing: he has uttered only one pan gesture, and two gestures for zooming, as displayed in table 6.2(a). The zoom gestures can clearly be decomposed into zooming in and zooming out. The hands of the subject pronate before zooming in (turning the palm of the hands away from each other), and supplinate before zooming in (turning the palms to each other).

Subject HMIG1002 has demonstrated four distinct pan gestures, which share identical motion. Table 6.2(b) shows that this subject has used four different hand shapes. The H hand was used in 75% of the pans, which was replaced by the G in 5 cases. It looks like if the subject did this to indicate more precise movement. The observations do not suggest any particular reason for the usage of the A and B hand. The zooms of subject HMIG1002 are very similar with those of subject HMIG1001. Subject HMIG1002 again uses different hand shapes for these signs, without any specific intended semantics.

Table 6.2: Final gestures

(a) HMIG1001

count	task	annotation
81	pan	Q/C/]{B5} Q/B5/& Q/B5/#{C}
12	zoom	S(Q/B/b Q/Bb/Z Q/B</#{C})
18	zoom	S(Q/B/a Q/Ba/Z Q/B>/#{C})

(b) HMIG1002

count	task	annotation
3	pan	Q/C/#{A} Q/A/& Q/A/]{C}
5	pan	Q/C/#{G} Q/G/& Q/G/]{C}
14	pan	Q/C/]{B} Q/B/& Q/B/#{C}
65	pan	Q/C/#{H} Q/H/& Q/H/]{C}
1	zoom	S(Q/C/>,]{G>} Q/G>/Z Q/G>/#{C})
6	zoom	S(Q/C/>,]{H>} Q/H>/Z Q/H>/#{C})
6	zoom	S(Q/C/>,]{B>} Q/B>/Z Q/B>/#{C})

6.6 Conclusion

The results of this experiment have shown that the Wizard of Oz paradigm suits quite well for obtaining gesture repertoires. Table 6.2 actually shows the extracted gestures from two subjects, which proves that the answer to the first question is a “yes”. It was believed beforehand that the human factor in this gesture recognition system would introduce significant latency in the interaction, however, both subjects have indicated that this was not noticeable. The operators were able to correctly track the subjects during both sessions, while both subjects have indicated to feel “in control”. As a matter of fact, both subjects were surprised by the immersiveness of the interaction.

The session has shown that using a variant of the ASL annotation language Stokoe the observations could be transcribed into a comparable form. Since the subjects showed both internal as external consistency, some ad hoc abstractions could be made, making it possible for future experiments to do a more efficient annotation.

Part II

The experiment

Chapter 7

Introduction

Part I addressed the question whether the *Wizard of Oz* paradigm was suitable for gesture prototyping. Moreover, it has developed a method to annotate on gestures, as well as an abstraction to reduce the annotation effort. While this previous part dealt with a small group ($n = 2$) of subjects, which were – by being involved with this research – strongly biased, this part aims at validating the results with a bigger group.

This part focuses on the development of a prototype gesture interaction for a selective set of tasks. Where traditional studies (see Bolt, 1980; Bowman and Hodges, 1997; Grossman et al., 2004; Vogel and Balakrishnan, 2005) have simply dictated the gestures for given tasks, this study turns things around by aiming at extracting the gestural repertoire from the user itself. The interaction will be "extracted" by a series of experiments in which the user is asked to solve certain tasks.

The previous part has focused on the details on how to conduct an experiment in which sets of gestures are delivered. This part has shown how to reduce large numbers of "similar" gestures into more generic "typical" gestures for certain tasks.

Again, this will be researched using the map application and experiment design of the previous session. The same map application will be used to induct interaction with the subjects, with some minor alterations. The assignments will be somewhat modified in order to reduce the time per subject and thus allow for more subjects to be "processed".

7.1 Research question

Unlike the previous session, where the form of the experiment was the centre of the research, this part focuses on the gestures itself. The main interest is what gestures people actually make using the map application, so the research question is formulated as:

> What typical gestures do people use for directing the map application?

Since in this part more and unbiased subjects are used, chances are that the proposed abstraction methodology of the previous part does not suffice, which could form a secondary challenge:

> Does the abstraction mechanisms as described in part (I) hold in this renewed experiment?

Chapter 8

Methodology

This chapter describes the methods used in the experiment. The experiment is setup exactly as in part I, enhanced with its recommendations. The session again loosely follows the "Wizard of Oz" paradigm, again with the subject knowing of this situation.

8.1 Application

The application will be enhanced with the recommendations of section 6.1:

- The (bit)map will be modified such that there is a big white area surrounding the visible map. This area is an easy means of providing visible feedback when the border of the map is reached;
- Additional pan and zoom paradigms will be implemented: pressing a key while moving the mouse will pan or zoom, depending which key is pressed;
- To solve the dual screen bug, the second screen of the workstations will be disabled.

8.2 Intrinsics

The setting of the experiment is exactly like the pretrial session, with some minor alterations, which are described in this section. First of all, the population of test subjects has been increased from two to ten. Although it is unfeasible to do a statistically sound experiment, a population of ten would increase the leverage of the results.

Moreover, the test subjects were significantly less biased than the pretrial subjects, which are close related to this research. The subjects of the previous parts were directly involved with the experiment and both are familiar with (multi touch) gesture interaction. All test subjects were given a fixed speech, in which the application, the two tasks (pan and zoom), the setting of the experiment but not the purpose was explained. By using written speech, which was read out, it was ensured that all subjects were treated equally.

The subjects were given less assignments, in order to fit all ten persons into one day of experiments. Each subject was scheduled into a time frame of fifteen minutes, as shown in table 8.1. After the introductory speech, the subjects were instructed to explore the interface of the application by playing around for 2 minutes. This exploration

phase was introduced to stimulate users to have their gesture set developed during the assignments, thus having more consistent gestures. In this phase, the subjects were encouraged to correct the operator when the application reacts unexpected to their actions.

Table 8.1: Time schedule per subject

minute	action
0	Pick up the subject from rendez-vous point
1	Signing of consent form, introductory speech
3	Start of "exploration"
5	First assignment
8	Second assignment
11	Third assignment
14	Escort subject back to rendez-vous point

8.3 Registration

The registration of the sessions differs somewhat from the previous session. Since the quality of the imagery was considered as (very) poor, even after enhancement, experimental IR-lighting was deployed. IR-lighting is known to get captured on video, while the human eye is not capable of seeing it. This potentially enhances the quality of the video images without reducing the visibility of the screen. The Motion Capture Lab, which was again used for the session, caters several IR-lamps as part of a Vicon setup.

Since it is only expected that this new lighting in the worst case delivers the same (poor) quality of images, but is not guaranteed, a second session with 10 "fresh" subjects will be scheduled. If the images of the IR-session prove usable, this second session will be cancelled.

Furthermore, apart from the video material, the subjects will be given a questionnaire, which will provide us with auxiliary information on the subjects. This auxiliary information can optionally help in interpreting the experiment results. This questionnaire is included as appendix B of this document and contains, besides personal information such as name, age, questions regarding the subject's affinity towards certain computer applications.

It is expected that affinity with the concepts described above will influence the gestural performance of the subject. People having extensive experience with applications like Google Earth or route planning software can have a certain bias towards those input patterns. Moreover, it is expected that people with rich computer experience will try to think of how gesture systems work, which could affect the gestures they employ. People with less computer experience are expected to consider the application more as a "black box". The first four of these concepts try to make a qualitative assessment of the computer knowledge of the subjects. Finally, people with a strong topographical knowledge of the area used in the map application will most certainly be able to solve the assignments, given a working interaction.

Section 6.1 suggests that the session could benefit from the operator sharing a view point with the camera. The problem is, however, that this implies that the operator would have to use a monitor connected to the video. This new view significantly alters

the session, thus it should be evaluated before applying it onto the new subjects. This evaluation requires a session like that described in part I, which was unfeasible due to the schedule of this research. It was chosen to not employ this altered view, in favour of the schedule.

8.4 Annotation

The first sessions suggested that in order to get a "typical" gesture set per task, the abstractions done after the annotation of the first session can be done before c.q. during the annotation itself. In this session, it will be tried to take advantage of this knowledge, reducing the (normally tremendous) amount of annotation work by moving these abstractions more towards the source of the annotation work flow. The same tooling (Anvil) as in the first session will be deployed.

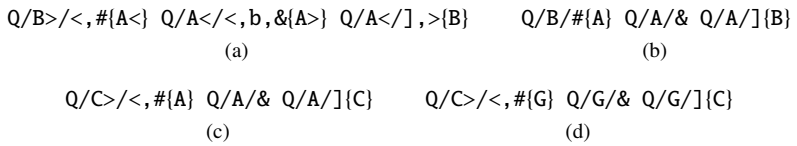


Figure 8.1: Several annotations

8.4.1 Abstraction

The abstraction will be done the same as in section 6.5, symbols without intended semantics will be discarded beforehand. Moreover, symbols which are implied by other motion symbols will also be left away. The symbol &, which is often used with pan movements, describes a motion in the pane parallel to the screen, which ubiquitously describes directional movement (>, <, ^ and v) and often pronation or supplination (a and b). These motion symbols will not be annotated when these motions are implied by the motion described by the symbol &. The same will be done for the motion in zoom tasks, described by the symbol Z.

This is exemplified by the annotations in figures 8.1(a) and 8.1(b). This movement illustrates a pan task implemented by the well-known "wave" gesture (see file `typical pan.avi`, in which the lower arm is moving from the dominant side of the body towards the centre by rotating the elbow, keeping the palm towards the "signee". The pronating (b) of the hand is a direct result, since in order to keep the palm facing the signee, the wrist has to compensate the rotation of the elbow. The movement start with a movement from the non dominant (>) to the dominating side (<) from neutral position. During the motion (&), it is inherent to the intended motion that the hand moves towards the dominant side (<). If one would leave away these redundant modifiers, the annotation can be cut down to its essence, as illustrated in figure 8.1(b). This latter annotation describes a base hand, closing up to a fist, which then moves freely in the pane parallel to the screen, which is followed by the hand opening up to a base hand.

On the other hand, figures 8.1(c) (`pan A.avi`) and 8.1(d) (`pan G.avi`) show two annotations of pan tasks, which are regarded as semantically different. Although both annotations show a hand closing from a neutral position indicating the "start cue" of the motion to be tracked, there is a fundamental semantic difference between the annotations. This key difference is the fact that 8.1(c) closes to a fist, where figure 8.1(d)

closes to a G-hand (see figure 6.1 on page 22). This difference in hand shapes is likely to have a cause from the semantic level, since the previous sessions have learnt that this G-hand is used to indicate a more fine-grained control of the motion. It is decided to be able to distinguish between these gestures, so these differences will not be abstracted from.

Furthermore, the video's will only be annotated on the following levels:

- Session;
- Assignment;
- Task;
- Gesture class.

The first three levels are already known from the first series of experiments, so do not require additional explanation. The "gesture phase" level was skipped for this experiment, since the previous sessions have learnt that this level does not contribute to the quality of the annotations. The fourth level, "gesture class" is introduced to replace the several Stokoe annotations from the previous sessions. The layer itself contains numbers, referring to a gesture class list (or map), which contains tuples (n, a) with n being a number and a an annotation in modified ASCII Stokoe. Instead of annotating each task, potentially repeating the same Stokoe annotations over and over again, the annotations will be drawn from this list. It is expected that this methodology increases the consistency of the annotations, since it reduces typographic errors in the annotations.

8.4.2 Work flow

This section describes the work flow of the optimised annotation process. In the first place, the video's will be recoded. At this moment, the segmentation on subject level will be performed as well, since the recoding tool (FFMPEG) provides an easy mechanism for this. Next, each video is sequentially loaded into ANVIL, to segment and annotate the other levels. Each video is segmented on the assignment level, giving each assignment the right label at the same time. The same will be done on the task level, which requires a quite precise segmentation, since the task segments are relative small, compared with the assignments.

The next layer will be pre filled with blank annotations using a small computer program, since the gesture class layer shares its segmentation with the task layer. This step saves considerable amounts of time and mouse clicks. In order to determine the gesture class, a separate list of gesture classes will be kept (see section 8.4.1). When there is no matching class, a new one will be created. This work flow will be performed in a breadth-first manner: first the complete annotation of the first level, then the first annotation of the second level, and so on.

8.5 Analysis

After the complete annotation of all video's, analysis needs to be done in order to answer the research question(s) of chapter 7. This analysis mainly consists of providing numeric statistics about the observed gestures of each subject. The average length of

each gesture per task per person will be determined, as well as the amount of gestures. Moreover, the occurrence count of each unique gesture will be set.

This statistical analysis will probably not fully answer the research question, since chances are that the abstraction methodology of the previous session does not apply to the new data of this session. If these abstraction techniques do not generalise the new data enough, new abstractions will be developed in order to provide the *typical* gestures of subjects for the given tasks.

Chapter 9

Results

This chapter deals with the results and observations of the experiment described in the previous chapter.

9.1 Registration

The experimental IR-lighting as proposed in chapter 8 demonstrated a huge increase in image quality. During the sessions, the IR-lighting provided a dim red glow for the human eyes, which did not hinder the subjects at all. The projected screen did not suffer from this extra light source. The videos quality enjoyed a huge improvement: the image was almost as clear as if filmed with bright daylight without any processing, so it was decided not to record a second session with traditional lighting.

9.2 Subjects

This section describes the sessions which each subject. Since the subjects had to fill in a questionnaire, each subject's answers will be given in a table. As described in chapter 8, the questionnaire consisted of three open questions (age, education and occupation), and two series of multiple choice questions. The first series multiple choice questions addressed the left or right handyness of the subjects, as well as the sex. The second series enquired the affinity of the subject with selected subjects, on a scale of 1 (few) to 3 (strong). Appendix C displays the results of this questionnaire.

Furthermore, some basic statistics about the observations are presented in table 9.1. Appendix D shows the Stokoe annotations of these subjects. The CD contains a folder named "`videos part 2`", containing all videos. The next subsections discuss the observations with the various subjects. As the observant reader may notice, there is no subject "HMIG2003", due to an error in the experiment scheduling.

9.2.1 HMIG2001

This subject demonstrated a complete different gesture implementation than what is observed from the previous sessions. Instead of the link between the displacement of the hand and map, this subject kept repeating the gesture for a task until the intended transformation of the view port was completed. This was observed for both tasks;

Table 9.1: Basic annotation statistics: time per assignment (t_n), average time per task (μ_t) and amount of tasks per assignment (n)

subject	t_{ass_1}	t_{ass_2}	t_{ass_3}	$\mu_{t_{pan}}$	$\mu_{t_{zoom}}$	n_{pan}	n_{zoom}
HMIG2001	0:48	0:36	0:47	0.85	2.10	42	17
HMIG2002	1:15	0:54	0:40	0.60	1.78	96	4
HMIG2004	2:31	1:20	0:42	2.25	1.33	51	74
HMIG2005	1:35	1:25	1:29	2.22	2.06	62	33
HMIG2006	0:39	0:22	0:27	1.45	1.62	37	12
HMIG2007	0:55	0:54	1:00	-	-	-	-
HMIG2008	0:56	0:55	0:17	1.56	1.68	27	20
HMIG2009	0:53	0:50	0:48	2.66	2.16	33	7
HMIG2010	1:00	0:44	1:07	2.67	3.10	35	9
HMIG2011	1:41	0:52	0:57	4.10	2.78	26	17
Average	1:13	0:53	0:49	1.78	1.83	45	21

zooming and panning. The subject purposely signed the gestures at the side of his body, optimising the view of the operator on this hand.

Subject HMIG2001 basically repeatedly points in the direction in which the map is intended to move, restricting himself to four directions: up, down, left and right. This pointing is done using an abducted arm, with the hand positioned next to the head. The subject is using his right arm as the dominant hand when doing this gesture, using its index finger to point in all directions, except for right. This exception for panning to the right is probably motivated by the fact that it is physically challenging to do this with the index finger.

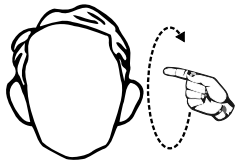


Figure 9.1: Zoom gesture of subject HMIG2001

The gesture for zooming, which is depicted in figure 9.1, can be described by a circular movement of the index finger in the plane orthogonal to the screen. This movement is displayed The direction (clockwise or anti-clockwise) indicated whether the level of zoom needed to be increased or decreased. Again, this gesture was repeated until the desired level of detail was reached.

The operator has indicated that these small, individual and fast pans are relatively hard to track. The average duration of an individual pan was 0.85 seconds, which is the second shortest average pan gesture observed in the whole series.

9.2.2 HMIG2002

The second subject employed an interaction scheme not unlike that of the first subject. Again, the typical pan movement consisted of a movement which was repeat until the desired view port transformation was complete. Similar to the previous subject, it was not very clear if the observed task was a series of short pans, or a gesture containing

a repetitive motion. The pan gestures were roughly the same of those of subject 2001, except that they were a bit shorter on average.

The zoom gestures of this subject can be described by a symmetrical movement with two hands, in which the palms face the head. When zooming in, the subject started with bent elbows, straightening these, moving the hands in a straight line from the body. The subject has never zoomed out.

Both operator and session host have have a strong suspicion that the interaction of the session was heavily influenced by the usage of speech by the subject. The subject employed vocal instructions with almost every gesture during the session. Such a verbal side channel is very hard to ignore for the operator.

9.2.3 HMIG2004

This subject demonstrated very similar interaction to the pretrial sessions. Moreover, the subject was extremely consistent, each task had only one typical gesture. The subject has indicated that it suffered from minor muscle pain due to excessive training the day before, which could have influenced his gesture performance. The operator has indicated that he had trouble keeping the zoom gestures separated, since these were implemented by roughly the same gesture. Zooming in was implemented by moving the hands to each other, using straight arms, starting with the hands removed from each other. Zooming out, on the other hand, was the inverse of this: the same gesture, started with the hands together, removing them from each other. Furthermore, the subject had difficulty with finding the target locations on the map.

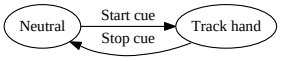


Figure 9.2: Gestural stages

9.2.4 HMIG2005

The subject indicated that he has seen several videos on the Internet in the field of multi touch interaction, such as Jeff Han's demo and the Mac book Air and that these videos probably influenced his gesture repertoire. The video's suggest that the subject interacts as if he is using a touchscreen. Somehow, it was not clear to the subject that the operator controlled the application, albeit this was disclosed in the speech. Although the pan and zoom gestures were quite like those of the pretrial sessions, the concept of the three stages (see figure 9.2) could not be applied to the observations. Since the operator had no trouble in detecting these stage boundaries and the video's do not cater any visible change in shape, it is hypothesised that these boundaries are marked by differences in speed.

9.2.5 HMIG2006

This subject demonstrated an extensive knowledge of the applied topography by finishing each assignment in a very short time. The pan movement of the subject was again similar to the observations of the pretrial session, although only one hand was exclusively used for panning. The other hand, however, was exclusively used for zooming.

This zooming was done using a flat (base) hand, next to the head. Moving the hand either up or down by bending the elbow, the subject zoomed in resp. out.

9.2.6 HMIG2007

This subject has shown a unique interaction: the absolute position of the hands marked the absolute "state" of the view port. During the whole session, the hand shapes remained constant, the only movement was the position of the hands in the pane parallel to the screen. The pan movement was implemented by one hand, which absolute position marked the absolute position of the map, which resulted in problems, since the arms of the subject were of limited length. This problem was solved by zooming in or out, positioning the map and zooming back to the original level. The zoom movement was done with the other hand, which was moved vertically to indicate the desired level of detail.

Since the movement of these gestures can only be described by the motion of the arms, which is not covered in (modified) ASCII Stokoe, it appeared impossible to make any useful annotations of this subject. Strictly spoken, the subject is constantly panning and zooming, even if there is no motion, which makes it impossible to distinguish between the tasks.

9.2.7 HMIG2008

The interaction of subject HMIG2008 is very similar to the pretrial observations, with no irregularities at all. There was some occasional confusion between zooming in and out.

From this subject on, the introductory speech was modified to include the fact that the operator "sees" the subject from the back and not through the camera in front of the screen. It appeared that the previous subjects assumed that the operator had a visual through this camera, which resulted in gestures "in front of the body", invisible for the operator.

9.2.8 HMIG2009

The difference between gestures and motion emphasis was not always very clear using only the imagery of the video. The audio of the video – the voice of the subject – provided key information to distinguish between these two. Similar to subject HMIG2005, the boundaries between the gesture stages were not applicable to this subject, which appeared to mark these with speed fluctuations.

The pan movement was similar to that of the pretrial observations, but without the change in hand shape to indicate transitions between the gesture stages. The zoom movement was implemented by a gesture in which the palm of the hand faces the signee and the hand is positioned next to the head of the signer. When the hand was moved towards or from the screen, the displacement of the hand indicated the intended change of level of detail of the map. At the end of the experiment, however, the subject has indicated that he would use the "two hand zoom" similar to the pretrial observations as well.

9.2.9 HMIG2010

This subject demonstrated "regular" pretrial-like gestures. He indicated to be familiar with multi touch interaction demo's, such as the Nintendo Wii. There was some occasionally confusion between zooming in and out, which were induced by a mis-

interpretation of the operator. The subject, however, quickly adapted to this altered interaction.

9.2.10 HMIG2011

Again, this subject showed "regular" pretrial-like gestures. This subject was familiar with video's of multi-touch interaction as well.

9.3 Conclusion

Except for subjects HMIG2001, HMIG2002 and HMIG2007, all subjects have provided suitable interaction to answer the research question by uttering consistent gestures. The gestures of subjects HMIG2001 and HMIG2002 require additional generalisation, since the previous abstraction techniques do not apply very well. Subject HMIG2007 appears to have produced gestures which is not annotatable in Stokoe. The next chapter will deal with the interpretation of these results and suggest an annotation technique for the first two subjects.

Chapter 10

Discussion

This chapter discusses and interprets the results as presented in the previous chapter. In the first place, the general experiment will be covered, followed by a discussion about the observed gestures. Furthermore, two sections are devoted to subjects HMIG2001 and HMIG2002, followed by suggestions for future research. Finally, this chapter will summarise this research by stating the conclusion.

10.1 General

Some of the subjects assumed that the operator's view on them was through the camera. A situation in which the camera would record the viewpoint of the operator would be better, since this is the closest to the semantic interpretation what can be recorded. This can be done by either moving the camera, or providing a monitor with the video feed to the operator.

The latter case has another benefit: if the operator exclusively uses this monitor, he or she cannot see the map. When the operator blindly controls the application, he or she cannot use its personal geographic knowledge to bias the experiment. It is expected that the fact that the operator could see the map has direct influence on the duration of an assignment.

One could think of an application in which the operator only sees a box on its screen, representing the view port on the map, like in figure 10.1. When the subject pans or zooms, the operator either moves the box or resizes it. In this case, the operator cannot "help" the subject by conveniently snapping the view port around a city, etc.

Moreover, the usage of speech can have the same effect on the experiment, since a lot of subjects (perhaps unconsciously) utter verbal queues such as "stop" and "yes, a little more to the left". Especially subject HMIG2002 appeared to rely heavily on this verbal side channel. During the design of these experiments it was more or less the idea that these little corrections were to be uttered using the gesture channel, since ultimately, some computer vision technique should replace the operator.

It is unclear whether these two "cheat opportunities" have any influence on the uttered gestures. It could be that they only reduce the amount of effort to solve an assignment, with the same gestures, but it could also be possible that new gestures would be uttered to make these corrections.

Besides from the aforementioned recommendations, the general setup of this experiment proved appropriate again. One subject even believed that it was a computer

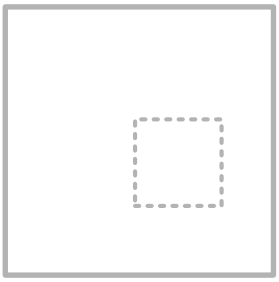


Figure 10.1: View port of the map. The dashed box is the view port, the bigger box is the map

instead of an operator interpreting its gestures.

Finally, it occurred several times that the operator accidentally confuses zooming in and out when a subject uses the zoom gestures as observed in the pretrial sessions. Although the gestures for zooming in and out could be reduced to one annotation, they could be regarded as each others inverse. It is still unclear what is the cause of this confusion, since the obvious metaphors for zooming (e.g., making something "bigger" with two hands) do not suggest an inverse interpretation.

10.2 Gestures

The individual pan gestures of subject's HMIG2001 en HMIG2002 can be considered as a cluster with one semantic. Both subjects kept repeating the gesture until the intended transformation of the view port was complete. This suggestion can be supported by the observation that both subject's had the shortest average pan gestures of all series.

As described in the previous chapter, subject HMIG2007 basically uttered only 1 gesture during the experiment. Besides the fact that this makes it impossible to segment the third level ("task"), the gestural motion cannot be fit in our dialect of Stokoe. This means that this subject proves that the employed annotation scheme is not "perfect". In order to annotate the motion of this subject, a severe extension of Stokoe would be needed, focusing on the motion of the arm. A language such as HamNoSys would probably cater for this.

All subjects, except HMIG2001, HMIG2002 and HMIG2007 have demonstrated "regular" pan gestures, except for a occasionally different hand shape. Two subjects, HMIG2005 and HMIG2009, appeared to use extremely subtle cues between the different gesture stages. As noted in the previous chapter, this could be indicated by differences in motion speed or very subtle motion. Although both operator and analyst have no trouble in distinguishing the stages, it was unclear what cues these stages.

Subjects HMIG2001, HMIG2002, HMIG2006 and HMIG2009 have introduced new zoom gestures, although subject HMIG2009 has indicated that he would use the "regular" zoom gesture the next time.

10.2.1 Clustering

The gestures demonstrated by subjects HMIG2001 and HMIG2002 suggest that the individual pan gestures making up a pan task could be clustered. This is supported by the fact that one individual pan gesture has no individual semantics besides being

part of the semantics of the whole series of pans. Moreover, the individual pans were very short ($\mu_{2001} = 0.85$ and $\mu_{2002} = 0.60$ which are both substantially lower than the average), making it very hard (if not impossible) to track the individual gestures for the operator.

These clusters of pans were defined by the criterion that all cluster members should have the same annotation and should follow the preceding pan immediately. A new annotation track was created with this clustered task layer, using Anvil's "snap" feature. Table 10.1 shows the statistics of the tasks before clustering and after clustering.

After clustering, it appears that the average length of a pan ($\mu_{t_{pan}}$) has decreased it deviation substantially. The distance of n_{pan} to the average, however, only decreases for subject HMIG2002. This could be explained by the fact that subject HMIG2002 uses more individual tasks per cluster and that subject HMIG2001 solved two of the three assignments in considerable less time. Unfortunately, the number of pans per subject cannot be used as an objective metric for gesture performance, since it depends on a wide variety of factors. In order to do something useful with this metric, it at least has to be normalised with respect to the geographical knowledge of a subject.

Table 10.1: Gesture statistics: without (top) and with (middle) clustering.

Subject	$\mu_{t_{pan}}$	$\mu_{t_{zoom}}$	n_{pan}	n_{zoom}
HMIG2001	0.85	2.10	42	17
HMIG2002	0.60	1.78	96	4
average	1.78	1.83	45	21
HMIG2001	1.69	2.10	22	17
HMIG2002	1.92	1.78	25	4
average	2.27	1.83	35	21

10.3 Abstraction

Analogue to the generalisation of the pan and zoom gestures in the previous section, the pan gestures of subjects HMIG2001 and HMIG2002 require some form of abstraction in order to provide a "typical gesture" for that task. All of subject HMIG2001's pans, for instance, are identical on semantic level. Three of the four groups of pans could easily be generalised by abstracting from the direction of the motion and orientation.

The problem is, however, that panning to the right includes an exception to the hand shape of the subject. In this context, we could allow the `A‘` and `G` hand shape to be annotated as the same symbol, e.g. `M`, indicating a pointing hand, with either thumb or index finger extended. This abstraction is motivated by the fact that in this context `A‘` and `G` are semantically identical and the subject would probably employ its `G` hand if it could to pan to the right. The abstraction from the direction could be implemented by the orientation symbol `d`, which, according to Stokoe's tradition, has different, but comparable meaning depending on if it is used for motion or orientation. When used as an orientation modifier, it indicates that the hand is pointing in a direction in the plane parallel to the screen. When used as an motion symbol, it indicates a dual motion, forth and back, in the plane parallel to the screen.

Table 10.2 displays the original and generalised gestures and statistics of subject HMIG2001.

When applying these generalisations on subject HMIG2002, who has similar pan movements as subject HMIG2001, four types of pan gestures remain, which is depicted

Table 10.2: Gesture annotations for subject HMIG2001, original (top) and generalised (bottom)

task	count	annotation
pan	26	Q/Gnv/v^
pan	10	Q/G>/><
pan	4	Q/A‘</<>
pan	2	Q/G^/^v
zoom	16	Q/G>/@
zoom	1	Q/G</@
pan	42	Q/Md/d
zoom	17	Q/G>/@

in table 10.3. The first two classes, which differ in the orientation of the hand shape, have equal meaning, but are used in different situations throughout the video. The subject uses the Q/Md/d gesture for the more course panning, while Q/Mf/d is used for the more fine grained panning when the “target” view port is getting close. There appeared, however, no specific reason for the usage of the two pan variants of lower frequency (Q/Bd/d and Q/B5f/d). Since these two account for only 8% of the uttered pans, they are considered as “noise”.

Table 10.3: Gesture annotations for subject HMIG2002, original (top) and generalised (bottom)

task	count	annotation
pan	35	Q/Gf/v^
pan	17	Q/Gnv/v^
pan	12	Q/G>/><
pan	10	Q/G^/^v
pan	7	Q/G</<>
pan	6	Q/G^/v^
pan	4	Q/B</<>
pan	2	Q/B5f/v^
pan	1	Q/B</><
pan	1	Q/B>/><
pan	1	Q/G>/<>
zoom	4	S(Q/Bf>/z,r)
pan	53	Q/Md/d
pan	35	Q/Mf/d
pan	6	Q/Bd/d
pan	2	Q/B5f/d
zoom	4	S(Q/Bf>/z,r)

The pans of the rest of the subjects (except for HMIG2007, of course) did not require any additional abstraction, since the abstraction of the pretrial sessions applied perfectly. All zooms did not result into any “diverse” annotations, so no additional abstraction was needed for these as well.

10.4 Research question

In chapter 7 the following question was identified as the research question:

> What typical gestures do people use for directing the map application?

Section 9.2 provides the most precise answer to this question, although this is a rather lengthy answer. The discussion as described by this chapter provides a more generic answer to this question. Since the map application basically consists of two tasks, the generic answer to the question is twofold: the typical gesture for panning and the typical gesture for zooming.

Statistics have shown that the most used gesture for panning is member of the "pretrial pan gestures", with which is meant a gesture with the following annotation:

$$\mathtt{Q}/S_1/M_1\{S_2\}\ \ \mathtt{Q}/S_2/\&\ \ \mathtt{Q}/S_2/M_2\{S_1\}$$

The symbols S_1 and S_2 represent hand shapes and M_1 and M_2 motion symbols. The motion symbols are each others inverses and either open or close the hand shape. The subject opens (or closes) the hand shape to indicate the start of the "tracking" of the hand, which is ended by the closing (or opening) of the hand shape. Hand shape S_1 is mostly covered by the symbol C, which was used by six of the eight subjects using this gesture for panning.

The gestures for zooming were slightly more diverse, although there is a most common gesture, which is also the only zoom gesture performed by multiple subjects. This zoom gesture is the generic version of the two-handed zoom, also observed before in the pretrial session:

$$\mathtt{S(}\ \ S_3/M_3\{S_4\}\ \ \mathtt{Q}/S_4/\mathtt{Z}\ \ \mathtt{Q}/S_4/M_4\{S_3\}\ \ \mathtt{)}$$

Again, S_3 and S_4 are hand shapes and M_3 and M_4 motion symbols. M_3 and M_4 are again each others inverses, opening or closing the hand shape. Also this time, M_3 and M_4 indicate the gesture phase transitions.

The other pan and zoom gestures were uttered by single subjects only, thus not regarded as "typical".

Section 9.2 also identifies another question:

> Does the abstraction mechanisms as described in part I hold in this renewed experiment?

It appeared that the abstraction mechanisms applied very well on the gestures of the subjects uttering the "pretrial paradigm", which means that the results are reproducible. The mechanisms did apply on some of the "new" gestures, but new rules had to be developed in order to produce the typical annotations, which proves that the abstractions prove to generalise in a way. This research question can thus be answered with a *yes*, with a side note that some additional abstractions were needed to cover *all* new gestures.

Chapter 11

Discussion

This chapter discusses the general results of this study as a whole. First, a general discussion on the results of the two parts will be done. Moreover, the research question will be answered, using these results. Finally, some recommendations for future research will be posed, based on the results and conclusions.

11.1 General discussion

All sessions in both parts did to some extent benefit from the verbal channel used by the subjects. Moreover, during the several abstractions and annotations the *intention* of the subjects was more or less guessed from the videos. These two facts could have been combined into some mechanism capturing the user's intention through speech. However, it could be challenging to capture an accurate view of the subject's intention in this way, since the subject should be explicitly aware of this intention and be able to formalise this into speech.

Several times in this study, it came forward that it could be beneficial that the operator shares a view with the recording camera. This idea was not implemented, since its success was not guaranteed, thus requiring an extra session to test this. This idea could have been implemented during the first part, but unfortunately did not came forward until the results.

This study so far only has been applied on a relative small group of not very diverse people (see appendix C). There were, for instance, only one woman involved in all trials, without any explicit reasons for this. An attempt could have been made to increase the leverage of the gestures, by using more diverse and probably more people in the second session. This was not done due to time constraints; the involved people were easy to contact and flexible in the logistical planning.

11.2 Research questions

The second part has shown that the observed gestures from the first part were observed in the larger, unbiased group of the second part as well. The second part has suggested that the gestural repertoire of subjects HMIG1001 and HMIG1002 are intuitive, since almost exact the same gestures are spontaneously uttered by unbiased test subjects. This is motivated by the definition of "intuition" according to Oxfords dictionary:

> the ability to understand or know something immediately, without conscious reasoning.

It is believed that when unbiased subjects spontaneously "invent" a certain gesture set, this could be called intuitive, according to this definition. Those gestures are extensively discussed in chapter 6.

There were, however, a few cases were this "intuition" did not apply. Subject HMIG2007, for instance, produced gestures which were not annotatable in the scheme, as developed in part I. Subjects HMIG2001 and HMIG2002 used gesturing that completely did not fit the intuition described above. This could be circumvented by allowing the intuitive gesture set to have different gestures for a single task.

The overall result of this study can be described as a success. The applicability of the Wizard of Oz paradigm to the gesture interface field for gesture extraction is proven, as well as a methodology for gesture analysis for this spectrum is given. The next section provide suggestions for future research on this topic.

11.3 Future research

This section will address several suggestions for future research. In the first place, a re-annotation using a different annotation scheme of the observations could be very useful, since Stokoe had to be extended in order to "fit" the data. This is further motivated by subject HMIG2007's gestures, which proved not annotatable in Stokoe. Furthermore, Stokoe does not cover any pure arm movement at all. Stokoe appeared to be *interpretable* as a more generic gesture interface annotation language, but this could be because the domain of this study (a simple map application) allowed this. A more generic language, which is not developed purely for (American) sign language, like for instance HamNoSys (see Prillwitz et al., 1989) should cover more generic gestures.

Next, it could be very interesting to discuss the search behaviour of the test subjects by looking at the way they try to reach their targets, which has probably its influence on the observed gestures. When a subject, for instance, is not sure where its target is, it typically "dwells around" or zooms out in order to find its target. During this phase, the gestures may be different than those used when the subject navigates directly to its target. Observations have shown that some subjects changed their hand shapes when doing more fine grained navigation, for instance. This search behaviour could be explored in a different experiment.

Another item being a good candidate for more research are the boundaries between the suggested gesture stages. A (very) typical pan gesture consists of a cue to signal the start of the tracking of the hand, the tracking phase and a stop cue. Most subjects have implemented very clear cues to indicate these boundaries, for example, by changing the hand shape. There are, however, a few subjects who provide extremely subtle cues, perfectly visible for the operator and video analyst, but it is very hard to determine of what they are made up. Research focusing on these subjects could provide interesting information for annotating these cues.

This research has so far only dealt with a series of objective measures of this kind of gesture interfaces. It would be interesting to see a subjective, qualitative analysis if this interface. This analysis could address the usability of this gesture paradigm.

In order to eliminate the influence of the operator's knowledge on the gesture performance, several measures could be taken in further experiments. As noted in section 10.1, the view of the operator could be altered such that he does not see the map it-

self. Moreover, the operator itself could be replaced by motion tracking and gesture recognition, operating on the gestures as provided by the subjects.

Appendix A

Annotation Manual

A.1 Definitions

This sections defines the key elements used in the annotation. Although some definitions may seem trivial, it is crucial that all annotators have a consistent interpretation to create a consistent annotation.

A.1.1 Segmentation

In the first place, the videos need to be segmented on several levels: session, assignment, task and transcription. The first level of segmentation places each session in a separate data file. Segments on the assignment level are positioned such that all tasks belonging to that assignment fit in the assignment block. The boundaries of the task blocks are chosen such that the whole of the gesture fits in the task. The boundaries of the assignment level is linked to the task level.

A.1.2 Stokoe

The language of choice is a adaptation of "Stokoe" (Stokoe et al., 1965), language designed to annotate American Sign Language using Latin characters. In Mandel (1993), Stokoe was adapted by replacing Latin characters by ASCII substitutes, to enable its use on computers. This adaptation is called "ASCII Stokoe". Besides the shift in character use, Mandel has enhanced the language by adding a few extra hand shapes, orientations and movements. Additional modifications, as described in section A.1.7. When using the term "Stokoe", we refer to modified ASCII Stokoe as described in this document. Furthermore, when not otherwise specified, we use Stokoe as described in Mandel (1993).

A.1.3 Structure

ASCII Stokoe's notation uses a location (always `Q` in this case), a hand shape, an orientation and motion. Orientation is considered as a modifier of the hand shape. The location, hand shape+orientation and motion are separated by a slash:

```
Q/B</>
```

In this example, the gesture is signed in the neutral position (Q, §A.1.4), with a base hand (B, §A.1.5) pointing towards the signer's non dominant side (<, §A.1.6), moving to the signer's dominant side (>, §A.1.6).

A.1.4 Position

Since Stokoe does not cover the variety of position observed in the first session, all gestures are "signed at" `Q`. Stokoe only covers locations on the body, and all observations were signed "in the air". `Q` is considered the "neutral position", used when the gesture is not signed on a specific location.

A.1.5 Hand shape

While annotating the sessions, these interpretations of (ASCII) Stokoe's hand shapes were used. NB: since these were the only observed hand shapes, the other hand shapes were not used. Refer to Mandel (1993); Stokoe et al. (1965); Wikipedia (2008) for their definitions.

A The closed hand or fist, as in ASL "a", "s" or "t".

B The "base" hand, a flat hand with the fingers straight ("b" or "4" in ASL).

B5 The base hand, with spread fingers, as in ASL "5".

C In "regular" Stokoe, the `C` refers to a "cupped hand", which is used slightly relaxed in our dialect, where it refers to a hand with (slightly) bent fingers.

G The pointing hand, as in ASL "1".

H Index and middle fingers together, as in ASL "h", "n" and "u".

A.1.6 Orientation and Motion

As mentioned before, in ASCII Stokoe, orientation is considered a modifier of the hand shape. Orientation symbols simply follow the hand shape, separated by a comma. An interesting property of Stokoe is the fact that motion and orientation share the same symbols. The symbol > means, when used in orientation, "orientated towards the non dominant side" and "moving towards the non dominant side" when used as a motion symbol.

v and ^ denote the hand pointing up or down, or moving up or down.

< and > respectively mean pointing or moving towards the non dominant and dominant side.

a and b stand respectively for pronation and supination.

t and f move or point towards and from the signer.

n bend the wrist, nod the hand.

close the hand.

] open the hand.

The motion symbols can be modified by a "end state", enclosed in curly brackets. This end state can optionally be used to describe the state of the hand at the end of the motion, when the end state is not trivial. We define the "end state" to be required when hand shape is altered by the motion. When the orientation is inherently altered by motion, but the hand shape stays the same, an end state is not required. The following example shows the usage of the "end state", showing a hand closing up to form the H hand.

```
Q/B/#{H}
```

Motion symbols can be compounded either sequentially, by simply using multiple characters or parallel, separating the characters with a comma. Orientation symbols can, of course, only be compounded parallel.

A.1.7 Additions

Since Stokoe was designed to transcribe *sign language*, it was no surprise that there were some aspects of the observations that could not be transcribed using Stokoe. The movement of the hand of both subjects during a pan or zoom task was often too complex to describe in Stokoe, requiring a lot of compounding. It was believed that this exact movement would not gain any knowledge required in this research, so two characters were introduced to describe these motions.

The symbol `&` indicates *some* movement in the plane parallel to the screen. The shape and orientation of the hand remain the same during this movement. This symbol is typically used to transcribe a pan task.

The symbol `Z` denotes *some* horizontal movement, parallel to the screen. Again, the shape and orientation of the hand remain the same. Zoom tasks were often annotated using this character.

Furthermore, to indicate that a gesture is performed by both hands, making it impossible to distinguish between a dominant and non dominant hand, the circumfix `S(...)` can be used. This circumfix is typically used annotating zoom gestures.

A.1.8 Gesture compounding

All zoom and pan gestures, as observed in the first sessions, can be decomposed into three stages. While this may resemble the well known gesture phases (McNeill, 1992), it is observed that these stages all occur during the stroke of the gesture. The second stage contains the semantics of these gestures: a movement which is meant to be mapped to the application. The first and last stage respectively signal the start and end of this movement.

The three stages are annotated as if they are uttered separately, as exemplified in figure A.1. Figure A.1(a) shows a typical pan gesture, where the hand opens and closes respectively in the first and third stage.

```
Q/B/]{B5} Q/B5/& Q/B5/#{B}   S(Q/B/f,> Q/B/Z Q/B/#{C})
```

(a) (b)

Figure A.1: Gestures with three stages

A.2 Guide

This section explains the practical "how to" of annotating a video.

A.2.1 Anvil

Since Anvil is the weapon of choice, a few pointers besides anvil's documentation are given here. In the pretrial sessions Anvil version 4.5 is used, so this version should be usable. First of all, since Anvil uses JMF as its media framework, JFFMPEG is required to ensure most common codecs are accepted. The alternative, FOBS4JMF is reported *not* to work in combination with Anvil. Take good care to follow the instructions of JFFMPEG, especially in registering the codecs through JMFSTUDIO.

A.2.2 Segmentation

The most efficient way to annotate a video is by creating a hierarchical annotation, starting on the top level ("assignment"). Next, each task (either "pan" or "zoom") needs to be segmented on the next level. Take good care that the whole task fits within the block. When two tasks follow each other very close, it may be wise to slow down the video (the scroll bar left to the video pane). Try to find the "middle" between the two tasks when the border is not very articulated.

Since the analysis of the pretrial learnt that the segmentation on the gesture phase level was not used to create annotations, this level can be skipped.

The next level is the actual annotation. The "blocks" on this level are linked to the task level in the Anvil specification, so there is no need to pay attention to the boundaries. In this phase, it is important to take a close look at the videos, preferably in slow motion. Use the definitions in ASCII Stokoe and the previous section[1] to find the best annotation.

A.2.3 Tips

This section describes a few tips from the experience of annotating the pretrial video's. These should increase the efficiency of the annotation process.

Empty annotations The Stokoe level can be annotated more efficiently when the "boxes" are created beforehand by editing the XML annotation. This is possible, since this level is linked to the task level in the XML annotation. When using some simple (shell) scripts, it is very easy to create a series of (matching) empty annotations.

Pause button One of the most efficient work flows in annotating the Stokoe level makes extensive use of the pause button. It may sound very trivial, but click pause before clicking on the button to end the annotation block. On the other hand, when the previous tip is taken in regard, the blocks are already ended.

Slow motion There is an option to slow down the video, the scroll bar is located directly to the left of the video pane.

Skip if unclear Since most tasks in the video are clear enough for annotation, there is not much spilled if an unclear annotation would to be skipped.

[1] In case of conflict, the previous chapter is leading.

Appendix B

Questionnaire

The questionnaire consisted of the following personal information:

- Sex;
- Age;
- Education;
- Occupation (employment, student, etc.);
- Left or right handed.

Furthermore, subjects were asked to indicate, on a scale from 1 (weak) to 3 (strong), to what extend he or she is affiliated with the following concepts:

- Computers in general;
- Computer games (FPS, shoot 'em up);
- CAD/CAM/DTP, graphic design tools;
- Internet;
- Map applications or route planners;
- The topography of Twente.

Appendix C

Questionnaire answers

Table C.1: Results of questionnaire

ID	HMIG2001	HMIG2002	HMIG2004	HMIG2005	HMIG2006
Age	32	34	24	19	36
Education	VWO	WO	MSc	VWO	VWO
Left or right handed	R	R	R	R	R
Sex	M	F	M	M	M
Occupation	R&D	Teacher	Student	Student	R&D
Computers	3	2	3	3	3
Computer games	2	1	2	1	1
CAD/CAM/DTP	2	2	2	2	2
Internet	3	3	3	3	3
Map applications	3	3	3	3	3
Topography of Twente	3	2	2	2	3

ID	HMIG2007	HMIG2008	HMIG2009	HMIG2010	HMIG2011
Age	26	25	21	20	28
Education	VWO	VWO	VWO	VWO	P
Left or right handed	R	R		R	R
Sex	M	M	M	M	M
Occupation	Programmer	Student	ICT	Student	Student
Computers	3	3	3	3	2
Computer games	1	1	2	2	1
CAD/CAM/DTP	1	2	1	1	2
Internet	3	3	3	3	3
Map applications	3	3	2	2	3
Topography of Twente	2	2	2	1	2

Appendix D

Gestures

Table D.1: Gesture annotations for subject HMIG2001

task	count	annotation
pan	26	Q/Gnv/v^
pan	10	Q/G>/><
pan	4	Q/A‘</<>
pan	2	Q/G^/^v
zoom	17	Q/G>/@

Table D.2: Gesture annotations for subject HMIG2002

task	count	annotation
pan	35	Q/Gf/v^
pan	17	Q/Gnv/v^
pan	12	Q/G>/><
pan	10	Q/G^/^v
pan	7	Q/G</<>
pan	6	Q/G^/v^
pan	4	Q/B</<>
pan	2	Q/B5f/v^
pan	1	Q/B</><
pan	1	Q/B>/><
pan	1	Q/G>/<>
zoom	4	S(Q/Bf>/z,r)

Table D.3: Gesture annotations for subject HMIG2004

task	count	annotation
pan	51	Q/B/#{A} Q/A/&/ Q/A/]{B}
zoom	73	S(Q/C/#{A} Q/A/Z Q/A/]{C})

Table D.4: Gesture annotations for subject HMIG2005

task	count	annotation
pan	62	`Q/C/],f{B} Q/B/& Q/B/#,t{C}`
zoom	33	`S( Q/C/],f{B} Q/B/Z Q/B/#,t{C} )`

Table D.5: Gesture annotations for subject HMIG2006

task	count	annotation
pan	37	`Q/Cˆ/]{Bˆ} Q/Bˆ/& Q/Bˆ/#{Cˆ}`
zoom	7	`Q/C</v,]{Bv} Q/Bv/ˆ{Bˆ}`
zoom	5	`Q/C</ˆ,]{Bˆ} Q/Bˆ/v{Bv}`

Table D.6: Gesture annotations for subject HMIG2008

task	count	annotation
pan	24	`Q/C/],f{B} Q/B/& Q/B/#,t{C}`
pan	3	`Q/C/#{G} Q/G/& Q/G/]{C}`
zoom	17	`S( Q/C/]{B} Q/B/Z Q/B/#{C} )`
zoom	3	`S( Q/C/#{G} Q/G/Z Q/G/]{C} )`

Table D.7: Gesture annotations for subject HMIG2009

task	count	annotation
pan	33	`Q/C/&`
zoom	5	`Q/C/f`
zoom	2	`Q/C/t`

Table D.8: Gesture annotations for subject HMIG2010

task	count	annotation
pan	35	`Q/C/],f{B} Q/B/& Q/B/#,t{C}`
zoom	7	`S( Q/C/b>,]{Bb>} Q/Bb>/Z Q/Bb>/#{C} )`
zoom	2	`S( Q/C/b>,#{Ab>} Q/Ab>/Z Q/Ab>/]{C} )`

Table D.9: Gesture annotations for subject HMIG2011

task	count	annotation
pan	25	`Q/C/],f{B} Q/B/& Q/B/#,t{C}`
pan	1	`Q/C/]{B5} Q/B5/& Q/B5/#{C}`
zoom	17	`S( Q/C/]{B} Q/B/Z Q/B/#{C} )`

Appendix E

CD contents

This report is accompanied with a CD-ROM, containing additional information relevant to this study. The CD is organised as follows:

```
README
videos/
   part1/
   part2/
code/
   viewer/
anvil/
   4.5/
   4.7/
data/
```

The `README` basically contains this appendix, the folder `videos` contains all video's of both parts, as well as general video's supporting parts of the text. The folder `code` contains the source code of the image viewer; the "map application". The folder `anvil` contains a small README on getting Anvil to run, as well as two versions of Anvil (which were both used). Finally, the folder `data` contains the anvil XML-files, as well as the AVI versions of the video's used in Anvil. The video's are encoded in XviD, a popular derivative of MPEG-4, which codec (ffdshow) is included in the folder video's.

Bibliography

R. A. Bolt. "put-that-there": Voice and gesture at the graphics interface. In *SIGGRAPH '80: Proceedings of the 7th annual conference on Computer graphics and interactive techniques*, pages 262–270, New York, NY, USA, 1980. ACM Press. doi: 10.1145/800250.807503.

D. A. Bowman and L. F. Hodges. An evaluation of techniques for grabbing and manipulating remote objects in immersive virtual environments. In *SI3D '97: Proceedings of the 1997 symposium on Interactive 3D graphics*, pages 35–ff., New York, NY, USA, 1997. ACM. ISBN 0-89791-884-3. doi: 10.1145/253284.253301.

T. Grossman, D. Wigdor, and R. Balakrishnan. Multi-finger gestural interaction with 3d volumetric displays. In *UIST '04: Proceedings of the 17th annual ACM symposium on User interface software and technology*, pages 61–70, New York, NY, USA, 2004. ACM. ISBN 1-58113-957-8. doi: 10.1145/1029632.1029644.

J. F. Kelley. An empirical methodology for writing user-friendly natural language computer applications. In *CHI '83: Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 193–196, New York, NY, USA, 1983a. ACM Press. ISBN 0-89791-121-0. doi: 10.1145/800045.801609.

J. F. Kelley. Six empirical steps for writing an easy-to-use computer application. Can be obtained from University Microfilms International; 300 North Zeeb Road; Ann Arbor, Michigan 48106, 1983b.

M. Kipp. Anvil - a generic annotation tool for multimodal dialogue. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, pages 1367–1370, 2001.

M. Kipp. *Gesture Generation by Imitation - From Human Behavior to Computer Character Animation*. PhD thesis, Saarland University, Saarbruecken, 2004.

M. Mandel. Ascii-stokoe notation: A computer-writeable transliteration system for stokoe notation of american sign language. Web, 1993. URL `http://www.speakeasy.org/~mamandel/ASCII-Stokoe.html`.

D. McNeill. *Hand and Mind: What Gestures Reveal about Thought*. UCP, Chicago, 1992.

S. C. W. Ong and S. Ranganath. Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(6): 873–891, 2005. ISSN 0162-8828. doi: 10.1109/tpami.2005.112.

S. Prillwitz, R. Leven, H. Zienert, T. Hanke, and J. Henning. *Hamburg Notation System for Sign Languages - An introductory guide*, volume 5. Signum, Hamburg, Germany, 1989. ISBN 3-927731-01-3.

W. Stokoe, D. Casterline, and C. Croneberg. *A Dictionary of American Sign Language on Linguistic Principles*. Linstok Press, Washington, D.C., 1965. ISBN 0-932130-01-1. reprinted by Linstok Press, Silver Spring, Maryland, 1976.

V. Sutton. Signwriting. online, February 2007. URL `http://www.signwriting.org`.

D. Vogel and R. Balakrishnan. Distant freehand pointing and clicking on very large, high resolution displays. In *UIST '05: Proceedings of the 18th annual ACM symposium on User interface software and technology*, pages 33–42, New York, NY, USA, 2005. ACM. ISBN 1-59593-271-2. doi: 10.1145/1095034.1095041.

Wikipedia. Stokoe notation - wikipedia, the free encyclopedia, 2008. URL `http://en.wikipedia.org/w/index.php?title=Stokoe_notation&oldid=183807989`. [Online; accessed 25-January-2008].

Acknowledgements

The author would like to thank Paul van der Vet and Dirk Heylen and especially Wim Fikkert for their support and feedback on this research. Moreover, Matthijs Bomhoff and Casper Joost Eyckelhof are thanked for their ideas, reviewing and brainstorms around this subject. Finally, thanks go out to Lukas van Schagen, Wieger Opmeer, Gerard van Bemmel, Koen Kooi, Kid Jansen, Joris Janssen, Michael Schrijver, Teun van Hemert, Annelies van der Veen, Lennard Klein and Maarten Aertsen for being a test subject.