Tangible Objects for the Acquisition of Multimodal Interaction Patterns

Ronnie Taib, Natalie Ruiz
National ICT Australia
Australian Technology Park
Eveleigh NSW 1430, Sydney, Australia
E-mail: {ronnie.taib, natalie.ruiz}@nicta.com.au

Abstract

Multimodal user interfaces offer more intuitive interaction for end-users, however, usually only through predefined input schemes. This paper describes a user experiment for multimodal interaction pattern identification, using head gesture and speech inputs for a 3D graph manipulation. We show that a direct mapping between head gestures and the 3D object predominates, however even for such a simple task inputs vary greatly between users, and do not exhibit any clustering pattern. Also, in spite of the high degree of expressiveness of linguistic modalities, speech commands in particular tend to use a limited vocabulary. We observed a common set of verb and adverb compounds in a majority of users. In conclusion, we recommend that multimodal user interfaces be individually customisable or adaptive to users' interaction preferences.

1. Introduction

Extensive research on speech and gesture based Multimodal User Interfaces (MMUI) began with Bolt (1980), more recently addressed semantic fusion of pen and speech input (Oviatt, DeAngeli & Kuhn, 1997) and untethered gesture recognition (Schapira & Sharma, 2001). However, the interaction design of such systems usually relies on ad-hoc techniques and pseudo-standards, resulting in arbitrary choices made by the user interface designer, and in the best case, User Centred Design (UCD) methodologies are applied iteratively to refine such choices based on end-user inputs.

Research by Taib & Ruiz (2005) suggested that the need for an expensive and time consuming electronic prototype can be avoided when the interaction to be modelled is applicable to tangible objects. A proposed method for acquiring a set of intuitive interface-function mappings from target users involves a human agent who interprets the user's input, and manipulates the tangible object accordingly. The output is observed by the user and can be corrected when a mistaken interpretation is made. The proposed methodology extends the traditional UCD framework, in particular, retains the main benefits of the Wizard of Oz (WoZ) approach, yet avoids the expensive development of electronic mock-ups and WoZ environment usually required in this process, e.g. (Bernsen, Dybkjaer, & Kiilerich, 2004).

In this paper, we propose a methodology of formative experiments for the design of multimodal interactions, especially when modalities are used in unusual combinations or contexts, for example using head gesture to manipulate virtual objects. We emphasise the need for individual customisation of multimodal interaction, by showing the diversity of user preferences even for simple tasks.

2. Methodology

2.1. 3D-Graph Navigation Application

The application domain selected is a multimodal interaction extension module for the GEOMI visualisation tool, allowing users to manipulate 3D-graph structures using speech and head gesture (Ahmed et al., 2005). However, the original multimodal interaction techniques were chosen by the developer in a relatively arbitrary fashion.

2.2. Experiment Design

An initial comparison between tangible objects and virtual (electronic mock-up) objects (see Figure 1) during the design phase, showed no major impact of tangible versus non-tangible on the users' behaviour, as perceived by the users themselves (Taib & Ruiz, 2005).

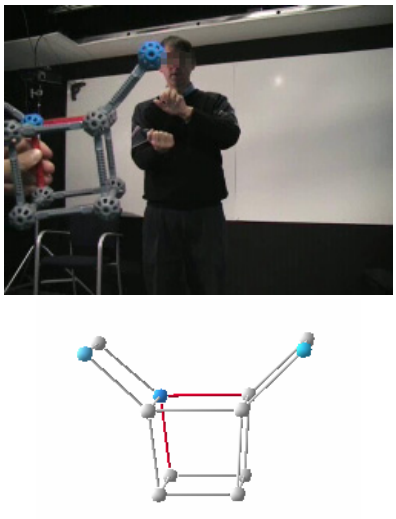

Figure 1: Subject interacting with the tangible object (top) and virtual representation (bottom)

In that experiment, 16 unpaid volunteer subjects (8 females and 8 males, aged 18 to 50) were instructed to teach a human "agent" how to interact with the system. Subjects were required to command object movements along all 6 degrees of freedom, in at least one direction. This process comprised the tasks of:

(a) mentally constructing an interaction pattern for each possible action in the target application, e.g. deciding that the tilting of the head to the right means the object should be rotated in the positive direction along the X axis;
(b) executing this pattern, e.g. the subject actually tilts her head to the right;
(c) the agent interpreting this pattern (e.g. he thinks this means a negative rotation along the X axis) and moves the object accordingly;
(d) the subject correcting the agent (e.g. using plain English) if the move did not correspond to her expectation built in step (a).

Subjects were videotaped from the front and side during the interactions. Four separate experiment conditions were administered, including speech, head and hands gesture separately and all in combination for object manipulations.

The video sequences were manually annotated, by the first author, listing any individual head gesture or utterance produced by each subject and the corresponding given object movement. We used the tangible form of the graph only as subjects have used the same behaviour with the tangible and virtual versions of the graph (Taib & Ruiz, 2005). We will refer to the tangible form of the graph as "the object" in the rest of this paper.

2.3. Hypothesis

In this paper we hypothesise that even for simple and seemingly straightforward multimodal commands, users will exhibit distinct preferences in terms of multimodal input patterns, highlighting the need for individually customisable or adaptive interfaces.

In case of head gesture, we propose the null hypothesis that users can be clustered by their gesture preferences, where for example, all users in a cluster would apply the same gesture to object move mappings *for all degrees of freedom*, leading to 2 distinct clusters of users. We further hypothesise that, what we will define as *direct mapping* may not be the preferred mapping for all users.

In case of speech input, we hypothesise that the vocabulary used may not span more than 5 to 10 words and that most combinations of verb and adverb may consistently express the same object moves across users.

2.4. Head Gesture Annotation

In order to generate objective annotations, an oriented 3D referential was used for both the head gesture and object movement descriptions, the origin being at the base of the subject's head and at the barycentre of the object respectively. The Y axis is directed towards the front camera and the Z axis is the upward vertical, as shown in Figure 2.

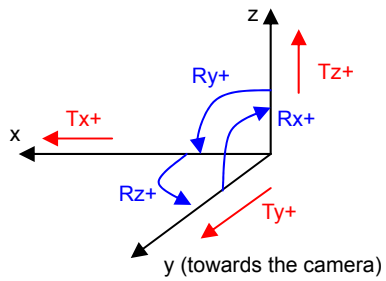

Figure 2: Movement Referential

As an example, we use Ry+ to refer to a rotation along the Y axis, in the positive direction (according to the oriented referential). We use Tx- to reflect a translation of the head or object "to the left" as seen from the subject's viewpoint.

Each head gesture is annotated with this referential, as well as the type of object move that was intended by that gesture. We asked subjects to limit their gestures to one direction along one degree of freedom at a time. However, subjects sometimes combined the primary movement with a secondary movement, e.g. slight rotation during a long translation, which we filtered out during data processing.

The last step of the data preparation was to ensure that one and only one mapping is used for each subject for any given degree of freedom and direction. We observed that subjects are always consistent with their own mappings (except for 1 input out of the total 144 head gestures annotated), so we simply keep only one record of each mapping when redundancies occur (e.g. a subject does the same gesture, intending the same object move, twice during the experiment).

Since all subjects did not perform moves in both directions for each degree of freedom, we assumed they would use the same type of gesture for the opposite direction, except that the direction of the gesture would be reversed. For example, if they gestured Rz- to move the object along Rz+, we can assume that they would gesture Rz+ to move the object along Rz-.

To support this assumption, we calculated that in 88% of the inputs, subjects only change the direction of their gesture to indicate a direction change for the object. This calculation is based on 50% of our data set that includes both directions for a given subject and degree of freedom. Finally, in the remaining 6% of the cases where the subject did not move the object along a specific degree of freedom at all, we exclude them from the average for this degree of freedom.

2.5. Speech Input Annotation

We did not impose any limitation on the possible speech inputs, in order to measure the span of syntax and semantics that subjects would use during the manipulation task. The subjects had various backgrounds ranging from engineering and technical to financial and

		Object move											
		Rx-	Rx+	Ry-	Ry+	Rz-	Rz+	Tx-	Tx+	Ty-	Ty+	Tz-	Tz+
Head gesture	Rx-	**53%**	33%							6%		**43%**	7%
	Rx+	47%	**67%**						7%	6%	6%	7%	**57%**
	Ry-			**100%**				7%					
	Ry+				**100%**				7%				
	Rz-					**67%**	33%		**40%**			7%	
	Rz+					33%	**67%**	40%					
	Tx-							**47%**	7%				
	Tx+							7%	**40%**				
	Ty-									**56%**	31%		
	Ty+									31%	**63%**		
	Tz-											36%	7%
	Tz+											7%	29%

Table 1: Cross-subjects head gestures, dominant one in bold

communications, hence the range of command styles is rather large. We analysed a total of 174 utterances produced by 15 subjects. We discarded all the spoken inputs of one subject because of their uncharacteristically high complexity and ambiguity. We classified the following elements in the collection of spoken inputs, so as to gain insight into their relationship to task difficulty:

- Verbs: Identification of mappings between specific verbs and specific object moves;
- Direction: Identification of expressions used to convey the direction of the degree of freedom and movement;
- References to the object: Identification of the vocabulary used to refer to the object or subparts of it;
- Dimension: Identification of how subjects limit the range or size of the move, either in relation to the object, absolute scales or relative to the subject's position.

3. Results

3.1. Head Gestures

Table 1 shows the preferences for each of the degrees of freedom of the object to manipulate between users.

The highlighted diagonal shows that, a *direct mapping* between the head gestures and object moves, e.g. head gesture is Rz- for a Rz- move of the object, is predominant for all degrees of freedom except the vertical translations Tz. The vertical translations tend to be achieved through a nodding type of move (Rx+/-) which is conflicting with the actual gesture for Rx moves of the object. Further analysis of the data showed that some subjects used a slightly different form of head rotation in these cases, generally wider or narrower. However, this is difficult to quantify and does not apply to the majority of subjects who simply used the exact same type of rotation to command a Rx and a Tz move of the object. Another effect strongly apparent in this table, is that for any given gesture along a given degree of freedom, in any one direction, there exists a gesture on the same degree of freedom, in the opposite direction, which we will call the *indirect mapping*. This is visualised by square blocks in the table, where the direct mapping is here in yellow and the indirect mapping in white cell background:

	Rx-	Rx+
Rx-	**53%**	33%
Rx+	47%	**67%**

The indirect mapping usually represents at least 30% of the preferred mapping, thus being a strong alternative. This supports part of the hypothesis that the direct mapping will not be predominant, however it shows that it is mainly for translations that this phenomenon applies.

In terms of clusters of multimodal interaction patterns, the most consistent cluster of subjects (see Figure 3) was of size 3 only (19% of the population). Interestingly, this cluster corresponds to direct mappings for all object moves. However, a minimum of 3 distinct mappings separates any other subject from this cluster and there is no other cluster of more than 2 subjects. This clearly rejects the null hypothesis that subjects consistently apply mappings across degrees of freedom. This results in a strong requirement for individual user adaptation in multimodal user interfaces.

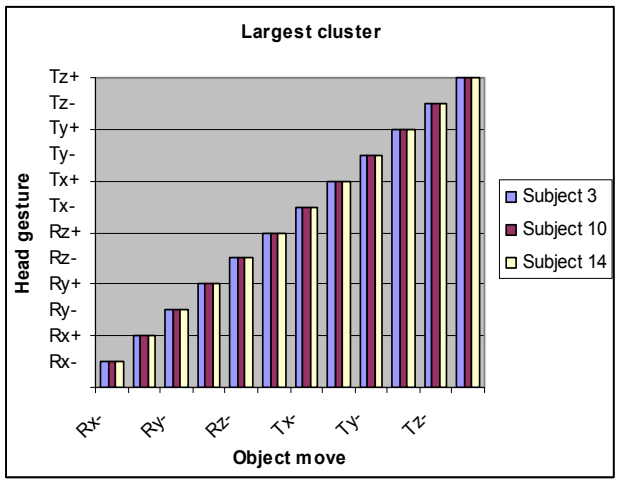

Figure 3: Largest cluster of subjects with identical mappings for all object moves

3.2. Speech Commands

Only 17 distinct verbs and 25 distinct direction adverbs or adverbial phrases were used inside the total 174 utterances. As can be expected, some verbs and adverbs are very popular while others are used rarely. We also observed that some of them have well defined and consistent meanings across users, while others were used for different commands across subjects and sometimes by the same subject. For generic verbs such as "rotate",we also analysed how much disambiguation could be achieved by using the combination of the verb and the adverb, but no syntactic or other semantic element, to denote commands.

Table 2 shows the most popular verbs and their most common meaning. While those verbs provide enough information to easily disambiguate between translations and rotations, only three of them enforce a specific axis and none a specific direction.

Verb	Dominant meaning	Occurrences	# of distinct subjects
Move	Translations	40 / 41	12
Rotate	Rotations	43 / 43	13
Turn	Rotations	16 / 16	8
Zoom	Ty (any direction)	9 / 9	6
Bring	Translations	7 / 7	4
Flip	Rx (any direction)	5 / 5	3
Spin	Rotations	11 / 11	3
Tilt	Rx (any direction)	4 / 4	3

Table 2: Eight most popular verbs and meanings
The number of occurrences with the dominant meaning over the total number of occurrences is in third column

Furthermore, the number of distinct subjects sharing the mappings between the verb and its dominant meaning shows that while three verbs are consistently used by 50% of the subjects, another 6 (not all shown in table) are used by 4 subjects or less and another 7 (not shown in table) are utilised by a single subject only..

The breakdown per adverb/adverbial phrases ranges even more widely, and only "up", "down" and "left" obtain remarkable popularity as well as consistent meanings. However, "right" is also popular but appears in several distinct rotation and translation commands, probably showing the limitations of the sample size used.

In terms of semantics, we observed that the adverbs usually fall into two categories of spatial references: an absolute positioning one, e.g. "clockwise"; and one relative to the user's position, e.g. "pointing to me". Some other referential parameters are sometimes used in the utterance, such as "in the horizontal plane", often in an ambiguous manner, e.g. "in the vertical plane" (not mentioning which specific vertical plane is considered) or "along the current axis". Those observations point out the need for either user training or disambiguation dialogues to palliate such shortcomings.

We also considered the compound (verb+adverb) as a means to identify commands more precisely.

Table 3 only shows the compounds with at least 4 occurrences in the data set. As initially hypothesised, some (verb + adverb) compounds are used consistently across users and without conflict with other commands. However, it mainly applies to translation commands, while rotation commands are usually overloaded to distinct or confounding meanings.

Compound	Number of occurrences	Distinct meanings	No of distinct subjects
Move left	10	Tx-	9
Rotate left	10	Ry-/+, Rz-/+	
Rotate right	9	Ry-/+, Rz+	
Move up	8	Tz+	8
Move towards me	6	Ty-	6
Rotate backwards	6	Rx-/+, Rz-/+	
<no verb> Up	5	Tz+	5
Turn right	5	Ry+, Rz+	
Zoom in	5	Ty-	4
<no verb> right	4	Ry+, Tx+	
Move down	4	Ry+, Tz-	
Rotate anticlockwise	4	Ry-/+, Rz-/+	
Rotate clockwise	4	Ry-/+	
Turn left	4	Ry-, Rz-/+	
Zoom out	4	Ty+	4

Table 3: Most popular verb + adverb compounds

3.3. Continuous vs. Discrete Moves

We discriminated between *discrete move* commands where the user expects the object to directly move to a given position, and *continuous move* commands where the subject expects the object to move until they stop their head gesture or speak a stop command.

As expected, head gestures are predominantly continuous with only 6% of discrete moves. The latter usually correspond to 90 degrees rotations, leaving the object symmetry axis aligned with the referential axis. Only one subject provided a 360 degrees move by rotating her whole body (though the usefulness of this move is questionable).

In terms of speech, 24% of the commands were discrete, as specifying angles and distances is simple via linguistic input. Such inputs fall into 4 categories:

- Angle/length specification, e.g. "15 degrees";
- Relative size, e.g. "one width of the object";
- Fuzzy quantitative input, e.g. "a little bit"
- Relative to the subject, e.g. "in front of me".

As an aside, we observed that only 38% of the spoken

commands explicitly refer to the object to move. It includes the words: “it” (20%) and “the object” (10%), as well as references to specific features of the object, e.g. “the front blue ball” (8%). These are used for some rotations when the subject has trouble finding a speech command that does not conflict with the previous ones.

4. Discussion

4.1. Speech vs. Head Gesture

This paper described the results of formative experiments targeting the initial steps of multimodal user interface design. The main result is the expansive range of multimodal inputs than can be expected from users when facing with a simple graph manipulation task. This paper presents only the results for two of the available of unimodal inputs channels elicited, namely speech and head gesture, but we expect to carry out the same methodology on the combined multimodal data collected during the experiment.

Head gesture was selected as an unusual but possible input modality for a 3D navigation application. We further constrained its expressiveness by limiting head gestures to 6 degrees of freedom, one at a time. Speech on the other hand is a very natural modality, even for the slightly unusual task of 3D navigation and positioning.

Even though head gesture was reported to be more challenging and less comfortable by the subjects, they were able to eventually construct consistent mappings between their gestures and the expected object moves.

Rotations over the Y axis attracted a 100% consistency across users, all of them tilting their head sideways to provide that command, as shown in Figure 4.

Figure 4: Example of Ry- gesture

The other rotations correspond to a *direct* or *indirect mapping* whereby the head movement replicates the expected move or its opposite direction combination.

Translations inspired a larger variety of mappings, not always dominated by the direct or indirect mapping. This highlights either some difficulty to physically perform the direct mapping, e.g. Tz+ requires people to move their head upwards from an already standing up position, or it may denote a preference for shorter, more discrete moves. On the other hand, speech operated rotations were found to be ambiguous, whereas many translations were intuitively done unambiguously yet consistently across subjects. These findings suggest the need to assess the combined multimodal (hands, head and speech inputs) condition, particularly focusing on whether subjects prefer head gestures to express rotations and speech to express translations.

4.2. Individual Customisation

Overall, the study results emphasise the need for user-specific customisation of multimodal interactions. It shows that there is no immediately accessible clustering of preferences across subjects even for a simple, finite set of inputs and tasks. Using predefined multimodal input patterns may result in moves totally opposed to the user’s expectations, since we showed that, in terms of head gesture as well as speech, the same input is often mapped to direct or indirect mapping according to the user. In fact, the same input may even target different degrees of freedom altogether, across users.

Furthermore, the exact type of gesture and its size may require some analysis towards customisation. Some users performed small amplitude, almost imperceptible gestures, whereas others used whole or upper body to perform wide movements as shown in Figure 5.

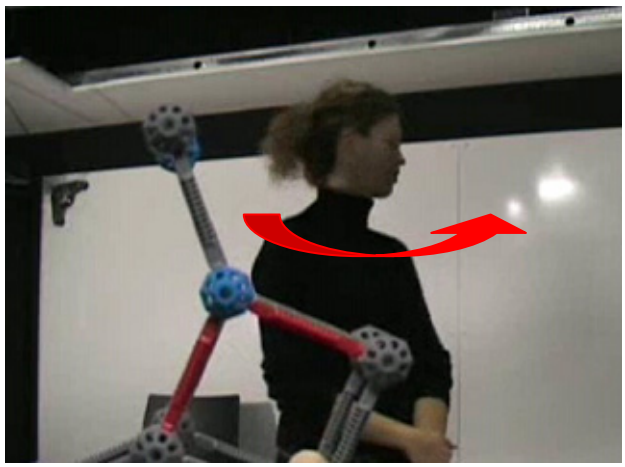

Figure 5: Large Rz+ gesture

4.3. Future Work

We collected hand gesture and combined multimodal data during the experiment that will require annotation and analysis. In particular, we will examine the occasions when people combine input modalities and how this relates to individual modality-based difficulties and limitations.

Hand gestures proved to be the most efficient way to manipulate the 3D object, given its affordances, as can be seen in Figure 1. We suggest investigating other application domains where tangible interfaces can be used to elicit multimodal interaction *without* access to direct

physical affordances. This relies on tangible reifications of abstract concepts or task, e.g. “what to do with an incoming email”.

Another suggestion is to extend the current work by conducting a longitudinal study assessing the performance and usability impact of predefined multimodal mappings that correspond or conflict with intuitive user preferences.

5. References

Ahmed, A. et al. (2005). GEOMI: GEOmetry for Maximum Insight. In GD05: 13th International Symposium on Graph Drawing. Limerick, Ireland: Springer-Verlag.

Bernsen, N. O., Dybkjaer, L. & Kiilerich, S. (2004). Evaluating Conversation with Hans Christian Andersen. In LREC'2004, Fourth International Conference on Language Resources and Evaluation (pp. 1011-1014). Lisbon, Portugal.

Bolt, R. A. (1980). "Put-That-There": Voice and Gesture at the Graphics Interface. In 7th annual conference on Computer Graphics and Interactive Techniques (pp. 262-270). Seattle, Washington, United States: ACM Press, New York, NY, USA.

Oviatt, S., DeAngeli, A. & Kuhn, K. Integration and Synchronization of Input Modes During Multimodal Human-Computer Interaction. In SIGCHI conference on Human factors in computing systems, (Atlanta, Georgia, United States, 22-27 March 1997). 1997, 415-422.

Schapira, E. & Sharma, R. (2001). Experimental Evaluation of Vision and Speech based Multimodal Interfaces. In PUI'01, Workshop on Perceptual User Interfaces (pp. 1-9). Orlando, Florida: ACM Press, New York, NY, USA.

Taib, R. & Ruiz, N. (2005). Evaluating Tangible Objects for Multimodal Interaction Design. In OZCHI'05, Annual Conference of the Australian Computer-Human Interaction Special Interest Group. Canberra, Australia: ACM Press.

6. Acknowledgements

The authors would like to thank the experiment volunteers for participating in the data collection.