A Corpus of Virtual Pointing Gestures

Ting Han
Bielefeld University

Spyros Kousidis
Bielefeld University

David Schlangen
Bielefeld University

firstname.lastname@uni-bielefeld.de

1 Introduction

Some things are better said, some things are better shown. When trying to convey the placement of objects relative to each other, one can use descriptions such as "the one is about two centimeters to the left of the other, and roughly one centimeter higher", or one can just place ones hands in a representation of this configuration and say something like "'one is here and the other one is here".

The type of gesture used in such displays has been called "abstract dexis" (McNeill et al., 1993) or "virtual pointing" (Kibrik, 2011), and it has been observed that these gestures have the remarkable effect of *creating* extralinguistic spatial referents for objects that are mentioned in the discourse, but are not in fact currently present. These referents can later even be re-used to form coreferential chains, as in the following example discussed by McNeill et al. (1993) (where square brackets mark the part of the utterance that is accompanied by the gesture described underneath the utterance):

(1) a. *and in fact a few minutes later we see [the artist]*
Points to left side of space.
b. *and uh she [looks over] Frank's shoulder at him*
Points to the left side of space again.

In this example, the first pointing gesture accompanies the first mention of the artist; in the second utterance, the pointing gesture accompanies the action and anticipates the object "at him" through reference to the location previously established as that of the artist.

Lascarides and Stone (2009) make the interesting proposal that such gestures do indeed call attention to a real location in shared space (which they denote with variables such as $\vec{p}$), but carry their semantic load via a mapping (v) into the conveyed location ($v(\vec{p})$) in the described situation, where the identity of the mapping is contextually determined. Configurations of locations indicated via such gestures (e.g. a $\vec{p_1}$ and a $\vec{p_2}$) then achieve their iconic value as a depiction of a configuration between the locations they are mapped into ($v(\vec{p_1}), v(\vec{p_2})$).

We were interested in how stable over time and how precise in their iconicity such mappings are in actual instances of use, with a view at how automatic understanding of such speech/gesture ensembles could be realized. We elicited and recorded multimodal spatial scene descriptions, and measured stability by looking at repeated gestural references, and precision by fitting a mapping between virtual referent locations and true object locations. We found that they can indeed be very stable throughout the course of a description (among 150 detected re-references, 81 of them are within 50 mm of the original references), and very accurately iconic. Moreover, we found a correlation between 'degree of iconicity' (that is, accuracy of the representation of the original configuration) and verbal effort.

2 The Corpus

In order to elicit pointing gestures in a virtual space, we designed a simple description task in which participants were shown an image on a computer screen for a brief time (10 seconds) and then were asked to describe it.

The images showed a configuration of four objects, and an arrow indicating a movement of one of the objects; this movement was also to be described. An example of such an image is shown in Figure 1. The objects were always simple geometric shapes, and at most two different colors were used. The scenes were designed in such a way that if gestures were used to indicate locations, this would have to be done successively (as there were

more objects than hands available to the subjects), and that for at the very least one object, namely the one that is to undergo the motion, there would be a need for a repeated reference.

In total, we recorded 311.63 minutes of video (by a HD camera) and motion capture data (by Leap motion[1]), of which 179.51 minutes contain speech. 14 participants took part in the experiment, each of them finished 29 scene descriptions on average (SD = 9.60). The analyses below were performed on 53 episodes (with 4 original references) from 8 dialogues, as not all data is annotated yet.

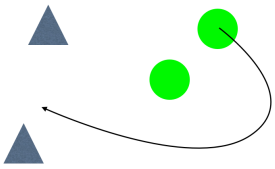


Figure 1: One of the scenes used in the experiments

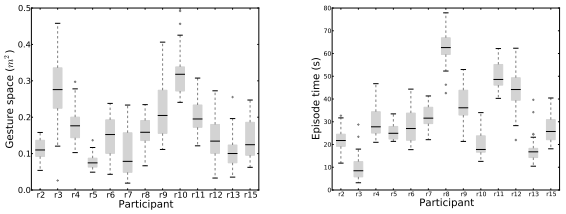
3 Results

We analyzed gesture space and episode length. As shown in Figure 2-(a), there was some variation both within and between subjects in terms of the size of the gesture space (calculated as the maximal area that their hands spanned during an episode). Figure 2-(b) shows that there is also variation in how long it took them to conclude episodes.

We used a shape matching method to compute how accurately the virtual pointing shape matched the original shape in the scene. Figure 2-(c) shows the histogram of matching errors. A matching error < 250 can be considered indicative of stable and precise gesturing performance.

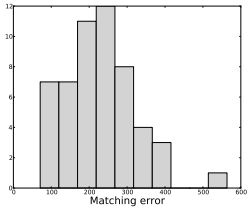
The re-reference precision is also analyzed. Figure 2-(d) shows the distance between the reference points (that is, a deictic gesture referring to the same object as a previous one) and the re-reference points. In the figure, we shift all the referent points to the original point (0, 0) and the black points stand for re-referents to the original point (0, 0). The x range and y range are the gesture space range in this example. We can see that although it's not quite precise when doing re-reference, but comparing to the whole gesture space, it's relatively small comparing to the gesture space. Figure 2-(f) shows the histogram of re-reference distance. Among 185 re-reference points, 161 of them are with re-reference distance < 150 mm, while gesture space is $900 * 671\, mm^2$.

The relationship between the number of words spoken in each episode and the corresponding gesture accuracy was also analyzed. Figure 2-(e) shows the result. We did linear regression, the correlation coefficient is 0.523. It suggests that when people gesture less accurately, they tend to need more verbal effort to describe the scenes.

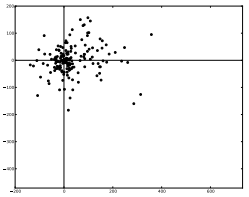


(a) Gesture space

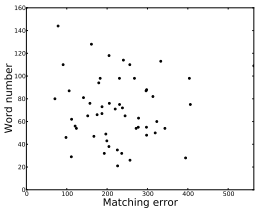
(b) Episode time



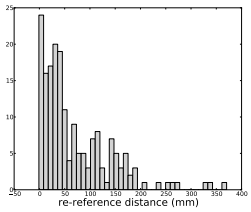
(c) Histogram of matching error



(d) Re-reference distance



(e) Word number and description accuracy



(f) Histogram of re-reference distance

Figure 2: Analysis results

References

Andrej A. Kibrik. 2011. *Reference in discourse*. Oxford University Press, Oxford, UK.

Alex Lascarides and Matthew Stone. 2009. A Formal Semantic Analysis of Gesture. *Journal of Semantics*, 26(4):393–449.

David McNeill, Justine Cassell, and Elena T. Levy. 1993. Abstract deixis. *Semiotica*, 95(1-2):5–20.

[1] www.leapmotion.com