Towards Automatic Understanding of 'Virtual Pointing' in Interaction

Ting Han
Bielefeld University

Spyros Kousidis
Bielefeld University

David Schlangen
Bielefeld University

firstname.lastname@uni-bielefeld.de

1 Introduction

When trying to convey, from memory, 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 these latter 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 in discourse be used to re-refer to the same entity; in our example, this could be done via "and *this* one [accompanied by pointing gesture] is".

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 precision by fitting a mapping between virtual referent locations and true object locations. We then used this mapping to retrieve from the set of all scenes the one that was being described. Using our matching method, we find that the gestures carry a good amount of spatial information for 45 out of 53 episodes. In current work, we are attempting to make this retrieval process incremental, and combine it with an understanding of the utterance that the gestures accompany.

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.

For all participants, the same series of 50 images was used that each is different from the others, but a time limit of 20 minutes was set for the whole experiment, and several participants did not complete the full set.

In total, we recorded 311.63 minutes of video (by a HD camera) and motion capture data (by Leap motion sensor[1]. Since we are interested in shapes in 2D, we only analyzed the data in x-y plane for all 3D data collected by Leap sensor.),

[1] www.leapmotion.com

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.

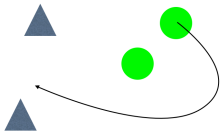

Figure 1: One of the scenes used in the experiments

3 Shape Matching and Scene Retrieval

Shape Matching The four objects in a scene form a shape with 4 vertexes, which can be represented as a matrix:

$$S = \begin{Bmatrix} x1 & y1 \\ \vdots & \vdots \\ x4 & y4 \end{Bmatrix} \quad (1)$$

in which rows correspond to object positions.

After getting the detected virtual pointing shape, we want to know how close it is to the original shape (S_o). However, due to different personal gesture space and pointing behaviors, the two shapes are not identical. We performed a shape matching method[2] to transform the detected shape to a target shape (S_t) which is most close to the original shape by shifting, rotating and scaling the detected shape, an example is shown in Fig 2.

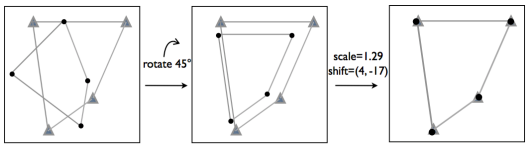

Figure 2: Shape matching

First of all, a randomly initialized transform parameter vector p is generalized:

$$p = [\theta, t_x, t_y, s] \quad (2)$$

where θ is the rotating angle; t_x and t_y stand for the shift value on x and y axis; s is the scaling parameter. For each row in matrix S we do rotation, shift and scaling with following equation:

$$S_t(x, y) = \begin{pmatrix} t_x \\ t_y \end{pmatrix} + s \begin{pmatrix} cos(\theta) & -sin(\theta) \\ sin(\theta) & cos(\theta) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \quad (3)$$

By minimizing the cost function:

$$E = min \parallel S_t - S_o \parallel \quad (4)$$

we get an optimized p which can transform the detected shape to the target shape S_t. We evaluate how close the detected shape is to the original shape with matching error, which is the distance between the target shape and the original shape.

Scene Retrieval We matched each detected virtual pointing shape with each of the 50 scenes that were prepared and ranked matching errors in ascending sequence. With good iconicity in the gestures, the matching error between virtual pointing shape and the original scene should have a low rank value, and the shape should pick out the scene that was actually described in the given episode from the set of all scenes that have been described.

4 Results and Discussion

For 21 of all episodes (39.62%), the gestured shape shows the smallest matching error among all candidates. In 30 episodes (56.60%), the gestured shape has error rank two. Consequently, the 2-best accuracy of using gesture shape information to retrieve the described scene is an impressive 96.22%. The remaining 2 episodes had a matching error above rank 2. A random selection baseline on this task would give an 1-best accuracy of 1.88%.

To fully evaluate these results, they would need to be weighted with a measure of similarity between the scenes that were to be described (because distinguishing between similar scenes based on spatial information is more difficult than between wildly different ones). But even in this form, the results already indicate that the gestures carry fairly accurate information about one aspect of the described scene. We take this as a starting point for our current work of combining this gestural information, in an incremental fashion, with information from the utterances that it accompanies (Kennington et al., 2013). The next step then will be to model recreation of the scenes from scratch, rather than selection from a set of candidates.

[2] http://glowingpython.blogspot.com/2013/06/shape-matching-experiments.html

References

Casey Kennington, Spyridon Kousidis, and David Schlangen. 2013. Interpreting situated dialogue utterances: an update model that uses speech, gaze, and gesture information. Proceedings of SIGdial 2013.

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.