A Corpus of Natural Multimodal Spatial Scene Descriptions

Ting Han, David Schlangen
Dialogue Systems Group // CITEC // Faculty of Linguistics and Literary Studies
Bielefeld University
firstname.lastname@uni-bielefeld.de

Abstract

We present a corpus of multimodal spatial descriptions, as commonly occurring in route giving tasks. Participants provided natural spatial scene descriptions with speech and abstract deictic/iconic hand gestures. The scenes were composed of simple geometric objects. While the language denotes object shape and visual properties (e.g., colour), the abstract deictic gestures "placed" objects in gesture space to denote spatial relations of objects. Only together with speech do these gestures receive defined meanings. Hence, the presented corpus goes beyond previous work on gestures in multimodal interfaces that either focusses on gestures with predefined meanings (multimodal commands) or provides hand motion data without accompanying speech. At the same time, the setting is more constrained than full human/human interaction, making the resulting data more amenable to computational analysis and more directly useable for learning natural computer interfaces. Our preliminary analysis results show that co-verbal deictic gestures in the corpus reflect spatial configurations of objects, and there are variations of gesture space and verbal descriptions. The provided verbal descriptions and hand motion data will enable modelling the interpretations of natural multimodal descriptions with machine learning methods, as well as other tasks such as generating natural multimodal spatial descriptions.

Keywords: Multimodal spatial descriptions, natural language, co-verbal gesture, abstract deictics

1. Introduction

When describing routes that are not visible in the situated environment, humans often accompany verbal descriptions with gestures to demonstrate relative spatial relations of landmarks or trajectories of the routes to follow (Emmorey et al., 2000; Alibali, 2005; Cassell et al., 2007). For example, when trying to help a person to locate a hotel not in current view, a description might be:

(1) You take the tram and get off at the stop "Schumacher street". Now here$_{[deictic]}$ is the tram station, here$_{[deictic]}$ is a fountain, if you walk$_{[iconic]}$ around it, you will see the hotel$_{[deictic]}$ on your left.

while the verbal description specifies the **landmarks** and **actions** (i.e., *tram station*, *walk around*), the deictic gestures encode the spatial layout of the landmarks with position information; the iconic gesture visualises the trajectory of the route. Only when combining speech and gestures together, it's possible to form a complete interpretation of the description, making it a challenging task even for human listeners (Schneider and Taylor, 1999).

In this paper, we present a corpus of multimodal spatial descriptions where hand gestures and speech are jointly used to describe spatial scenes. The corpus includes data collected from two experiments, a scene description experiment and a spatial description experiment. In the two experiments, participants received different instructions to perform the task and got different feedback signals when performing the task (see Section 3.1. and Section 4.1. for details). The former experiment focused on eliciting intuitive multimodal descriptions, while the latter experiment aimed to elicit spatial descriptions with human-computer interaction oriented instructions and constrained gesture space.

In the **scene description experiment**, we aimed to collect intuitive multimodal descriptions. Participants were given a spatial scene description task without instructions on how to perform the task. That is, they described intuitively, either only using speech or using both speech and gestures. The results show that participants often intuitively use gestures in such spatial descriptions and the deictic gestures reflect spatial layouts of landmarks. However, the varied gesture spaces and relatively limited tracking space of existing devices often make it difficult to track hand motion. Hence, we designed the **spatial description experiment** with a somewhat more constrained setup.

In the spatial description experiment, participants were told that they were describing to a computer program (WOz setting; Kelley (1983)). They were suggested to use gestures and restrain their hand gestures in the effective tracking area, so that the computer can "see" the gestures and understand the descriptions better. This setup resulted in a dataset with sufficient hand motion data, while none of the participants reported unnatural gestures due to the limited gesture space.

We have made the following resources of the corpus publicly available: scene information which were used to elicit the descriptions, the transcriptions of speech, recorded hand motion data, and annotations of deictic gestures and speech. (Han et al., 2018) modelled real-time understanding of spatial descriptions using the data of the spatial description experiment. The results show that incorporating hand gestures not only leads to more accurate interpretation of such descriptions, but also leads to earlier final correct interpretations.

2. Related work

Easily available video and audio recording devices have facilitated conversational/discourse level analysis of speech-gesture communications (Lücking et al., 2010; Quek et al., 2002; Schiel et al., 2002). Although these corpora provide natural multimodal communications and detailed annotations, it's a difficult research problem in itself to extract

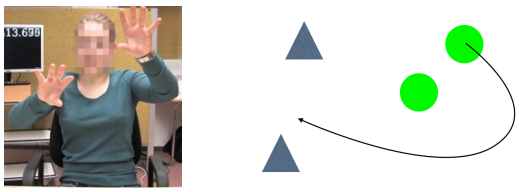


Figure 1: Providing a scene (*right*) description with speech and gestures (*left*). The arrow indicates the movement of the green ball.

3-D gesture features from these videos. Motion tracking sensors that recently have become readily available as well (e.g., Kinect[1] and Leap sensor[2]) make it possible to record large scale 3-D gesture datasets, such as (Tompson et al., 2014; Marin et al., 2014; Liu and Shao, 2013; Sadeghipour and Morency, 2011) and datasets mentioned in (Cheng et al., 2016); however, most of these existing datasets are collected for gesture classification tasks without accompanied speech. (Fotinea et al., 2016) presented a dataset of multimodal commands, where gestures and accompanied are both recorded. However, the gestures are with defined meanings that are independent of speech. In addition to previous datasets, we present a corpus composed of natural multimodal communications with high-resolution hand motion data, in which the meaning of gestures depends on accompanied speech.

3. The Scene Description Experiment

In this experiment, we aimed to collect intuitive scene descriptions. Participants were shown simple scenes (as shown in Figure 1) briefly and asked to describe the scenes from memory. There were no instructions on how to perform the task, hence participants described intuitively, either describing with speech (*mono-modal*) or with speech and gestures (*multi-modal*).

3.1. Task design

We designed a simple scene description task to elicit natural scene descriptions. Participants were asked to describe scenes composed of four simple objects and an arrow which indicates the movement of the object (as shown in Figure 1), intended to trigger deictic and iconic gestures.
We generated 50 such scenes. In each scene, the four objects are with two colours and two shapes. Object *colour*, *shape*, *size* and *position* were randomly selected when the scenes were generated. The arrows originate from one of the objects and point to somewhere near another object. To accurately describe the movement and the spatial configurations after the movement, participants will need to demonstrate the spatial layout with gestures as spatial configurations are difficult to convey with natural language.
To investigate the natural behaviours of such descriptions, participants were only asked to describe the spatial configurations of the objects and the movement indicated by the arrow. Gestures were not mentioned in the instruction.

[1] `https://developer.microsoft.com/en-us/windows/kinect`

[2] `http://www.leapmotion.com`

To elicit accurate descriptions, participants were told that another person will watch the descriptions later and try to recreate the scenes. Describing accurately will make the re-creation task easier for the other person. For each description, the scenes were briefly (10 seconds) shown on a computer screen. After the scene disappeared, participants started to describe.
Each participant described for 20 minutes. In total, 15 participants (native German speakers; students from Bielefeld University) took part in the experiment.

3.2. Recording setup

We recorded audio and video with a HD camera. The hand motion was tracked with a Leap sensor, a portable device composed of two monochromatic cameras and three LED infrared sensors. The hand motion data was recorded with MINT Tools (Kousidis et al., 2013). Both videos and hand motion data were recorded with timestamps.
In the experiment, participants were seated in front of a table. Right across the table and in front of the participant is a HD camera to record audio and videos. A Leap sensor was placed on the table in front of the participant.
On the right side of the table is a monitor which displays the scenes. An experimenter was seated next to the participant to display the scenes. For each scene description, the experimenter clicked a button to show the scene for 10 seconds, then turned the screen to black. After that, participants started to describe. When the description ended, the experimenter advanced to the next scene.
The Leap sensor tracks hand movements and outputs data frames to represent hand motions as following:[3]

- **FrameID**: integer, a unique ID assigned to this data frame.
- **hand number**: integer, the number of tracked hands.
- **hand confidence**: float, ranging from 0 to 1. It indicates how well the internal hand model fits the observed data.
- **hand direction**: 3-D vector. The direction from the palm position toward the fingers.
- **hand sphere centre**: 3-D vector. The centre of a sphere fit to the curvature of this hand.
- **sphere radius**: float, the radius of a sphere fit to the curvature of this hand.
- **palm width**: float, the average width of the hand (not including fingers or thumb).
- **palm position**: 3-D vector, the centre position of the palm in millimetres from the Leap Motion Controller origin.
- **palm direction**: 3-D vector. The direction from the palm position toward the fingers.

[3] For detailed descriptions of these features, please refer to the official SDK manual`https://developer.leapmotion.com/documentation/python/index.html`

- **palm velocity**: 3-D vector, the rate of change of the palm position in millimetres/second.
- **finger length**: float, the apparent length of a finger.
- **finger width**: float, the average width of a finger.
- **joint direction**: 3-D vector, the current pointing direction vector.
- **pinch strength**: float, the strength of a pinch pose between the thumb and the closest finger tip as a value in the range $[0, 1]$.
- **grab strength**: float, the strength of a grab hand pose as a value in the range $[0, 1]$. 0 when the hand is open. As a hand closes into a fist, the grab strength increases to 1.
- **finger type**: integer, the integer code representing the finger name. 0 for thumb, 1 for index, 2 for middle, 3 for ring, 4 for pinky.

3.3. Data processing

A sample description is shown as follows:

(2) a) Hier$_{[deixis]}$ ist ein graues Dreieck und hier$_{[deixis]}$ ist ein grüner Kreis hier$_{[deixis]}$ ist noch ein grüner Kreis und hier$_{[deixis]}$ ist noch ein graues Dreieck und von$_{[iconic_start]}$ dem oberen grünen Kreis geht rechts neben dem anderen grünen Kreis$_{[iconic_end]}$ zwischen den beiden Dreiecken nach links ein Pfeil.

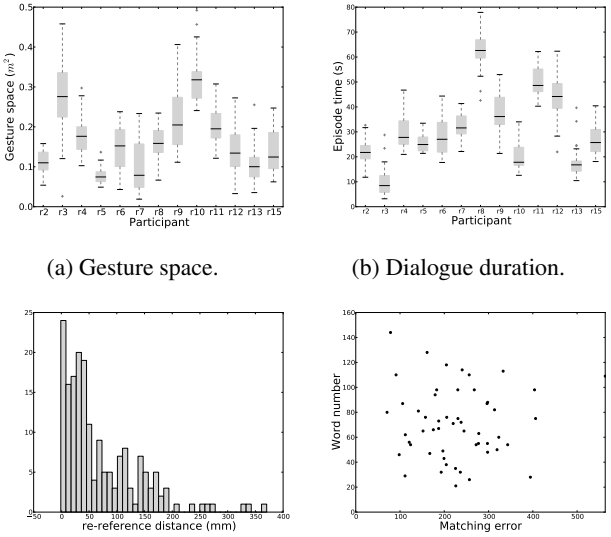
b) *Here$_{[deixis]}$ is a grey triangle and here$_{[deixis]}$ is a green circle here$_{[deixis]}$ is another green circle and here is another grey triangle and from$_{[iconic_start]}$ the upper green circle goes right next to the another green circle$_{[iconic_end]}$ between the two triangles to the left, the arrow.*

Transcription The audio was manually transcribed by native speakers. The transcriptions were temporally aligned with the audio and video recordings on the word-by-word level using an automatic forced alignment approach. We annotated each scene description with corresponding **scene ID** by watching the recordings in ELAN.[4] Each scene description was segmented into individual object descriptions which were annotated with corresponding **object ID**. For instance, the scene description in Example (2) was annotated as *Scene 15*, while the object description "*here$_{[deixis]}$ is a grey triangle*" was annotated as *object 1*.

Gesture annotation Conventionally, each deictic gesture is divided into several gesture phases: pre-stroke, stroke, stroke hold and retraction (Kendon, 1980). During the stroke hold phase, hands stay in the gesture space to indicate object positions, hence, it's the most informative phase. We manually annotated the stroke hold phase of each deictic gesture. The annotation was done by watching the video recordings and the described scenes using ELAN. Similar to natural language annotations, we labeled the *stroke hold* phases with the *object ID* of referential objects.

With the recorded timestamps, hand motion data was aligned with video recordings. Accordingly, the hand motion frames were labeled as *stroke hold* frames or *non-stroke hold* frames according to the timestamps.

As aforementioned, iconic gestures were also involved in the descriptions. For instance, in (2), while describing the movement of the grey triangle with utterance "from the upper green circle goes right next to another green circle", the participant drew a line in the gesture space to indicate the trajectory of the movement. We annotated the start and end iconic gestures with $[iconic\_start]$ and $[iconic\_end]$.



(a) Gesture space.

(b) Dialogue duration.

(c) Histogram of re-referential distance.

(d) Word number and description accuracy.

Figure 2: Preliminary analysis results.

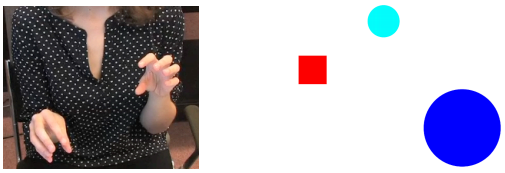
3.4. Preliminary analysis

Varied gesture spaces We calculated the maximal area that each participant's hands spanned during all their descriptions as their gesture space. As shown in Figure 2a, there are variations both within and between subjects in terms of the size of the gesture space which make hand motion tracking and gesture interpretation challenging tasks.

Referential accuracy We also analysed the re-reference accuracy. Figure 2c shows statistics of the reference distance between a deictic gesture and its original gesture. Among 185 re-reference points, 161 of them are with re-reference distance < 150 mm, while the maximum gesture space is $900 \times 671\, mm^2$.

Gesture accuracy of spatial configuration We used a shape matching method to compute the distance between the configurations of gestures and corresponding object positions. The distance was used as a measure of gesture accuracy and compared to the words spoken in each episode, as shown in Figure 2d. The result suggests that when people gesture less accurately, they tend to need more verbal effort to describe the scenes. We did linear regression to analyse the relationship between the number of words spoken in

[4] https://tla.mpi.nl/tools/tla-tools/elan



(a) (b)

Figure 3: Providing a spatial description with speech and gestures: *here is a red square, here is a light blue circle ...*

each episode and the corresponding gesture accuracy. 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.
While the annotation work of iconic gestures is still going on, we leave it as future work to analyse statistics of iconic gestures.

4. The Spatial Description Experiment

In this experiment, we focused on collecting multimodal spatial descriptions with tracked hand motion data which enables the modelling of such multimodal behaviours with machine learning methods. Therefore, we simplified the description task by removing the iconic element (i.e., the arrow) from the scenes and suggested participants to describe with speech and gestures.

4.1. Task design

To elicit multimodal spatial descriptions, we simplified the description task. Instead of describing configurations of four objects and a movement, participants were asked to describe scenes only composed of two circles and one square, as shown in Figure 3. To further reduce the cognitive load of scene memorising, we displayed the scenes on the screen throughout descriptions.
We generated 100 such scenes. The colour and shape of each object were randomly selected when the scenes were generated. There were 6 colours and 2 shapes (*square, circle*). Each of them had the same chance to be assigned to an object. The size and position of each object was randomly generated. The object size ranges from 0.05 to 0.5 in ratio to the size of the scene image. The object positions were adjusted until none of the objects overlap with each other.
Participants were told that they will describe scenes to a *computer program*. The computer will try to understand the descriptions by listening to the verbal descriptions and watching their hand gestures. After each description, the computer displays a score on the screen which ranges from 1 (worst) to 5 (best) and indicates how well the computer understands the description. In reality, the score was from the experimenter who rates the descriptions according to the number of mentioned object attributes.

4.2. Recording setup

The technical recording setup is similar to previous experiment, except that the **hand types** (left or right) were also recorded with a new Leap SDK (SDK v2.3.1). We also placed a monitor in front of the participant to display their hand motions, encouraging them to gesture in the effective tracking area.
At the beginning of the experiment, the experimenter first introduced the task and all the recording devices to the participants, then demonstrated a description with speech and gestures. Participants were suggested to describe with speech and gestures and mention *shape, size, colour* and *relative positions* of the objects. They also had several minutes to play with the Leap sensor to get familiar with the effective tracking area of the sensor.
In total, 13 participants (native German speakers) took part in the experiment (None of them took part in the previous experiment). Each of them described for 20 minutes.

4.3. Data processing

The data was processed and annotated in the same way as previous experiment. A sample description is shown as follows:

(3) a) Hier$_{[deixis]}$ ist ein kleines Quadrat, in rot, hier$_{[deixis]}$ ist ein hellblauer kleiner Kreis und hier$_{[deixis]}$ ist ein blauer grosser Kreis.

b) *Here*$_{[deixis]}$ *is a small square, red, here*$_{[deixis]}$ *is a light blue small circle and here*$_{[deixis]}$ *is a blue big circle.*

Scene representation We represented each scene as a composition of three objects. Each object was represented with 4 attributes: *colour*, *shape*, *size* and *position*. For example, a real valued position coordinates can be represented as $x : 0.1, y : 0.2$. The position was further discreted into *top, middle, bottom* vertically and *left, middle, right* horizontally. So that with the scene ID and object ID in each multimodal description, corresponding object attributes can be retrieved to reconstruct the described scene. For example, the pink circle in Figure 1 is represented as following:

- SceneID: *scene 1*
  - Object ID: *object 1*
  - Colour: *red*
  - Shape: *circle*
  - Coordinates: {*x: 0.22, y: 0.54*}
  - Horizontal position: *left*
  - Vertical position: *middle*

4.4. Preliminary analysis

Varied verbal descriptions Although participants were suggested to encode *colour*, *shape*, *size* and *relative positions* of objects in the descriptions, they were allowed to form descriptions in their own way. The collected data also reflects varied verbal descriptions. For example, the same colour was described with various expressions. *Pink* was also described as *lila*. *Cyan* was sometimes referred as *light blue*. *Circles* were referred as *circle* or *ball*. The vocabulary size of the corpus is 291.

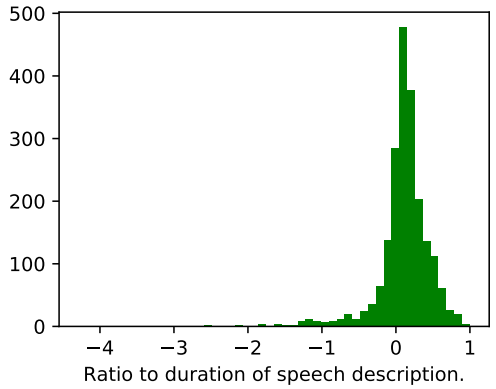


Figure 4: Temporal relations between speech and deictics.

Varied gestural behaviours From the data, we observed that when describing, sometimes participants use one hand each time to demonstrate the object position in the gesture space, hence, the listener needs to keep track of previous object positions to form a whole mental representation. Alternatively, some participants demonstrate with two hands in the gesture space to show relative positions. Among 830 description episodes, 637 descriptions (76.7%) involved the use of both hands; 193 (23.3%) with one hand. In both cases, the hand gestures convey spatial layout of the objects.

Temporal relations of speech and gestures Speech and co-verbal gestures are in parallel, and bear close temporal relations between each other (Ragsdale and Fry Silvia, 1982). We analysed the temporal relations of start timings between speech and gestures, as shown in Figure 4. Among 2074 speech-deictic ensembles, 24.5% deictics precede accompanied verbal description; 47.3% deictics occur in the first quarter of verbal descriptions. The parallel characteristics could benefit multimodal interpretation tasks on the incremental level (Han et al., 2018).

Indicating shape/size with deictics Deictic gestures have been extensively studied for positional information. However, humans often encode more than positional information while "pointing". In the collected data, we observed that, beside positional information, participants also encode shape and size information in gestures. For instance, some participants used different hand shapes when referring to circles and squares. Moreover, when mentioning objects with larger sizes, they tend to form larger hand spheres. This suggests that in future work, gesture interpretations should consider various dimensions of the information.

5. Availability

The hand motion data, anonymised transcriptions and annotations of the second experiment are publicly available[5] under the ODC Public Domain Dedication and Licence (PDDL).[6] To access the audio and video recordings, please contact the authors. Instructions on how to use the data are also available `https://tingh.github.io/resources/scene_description`.

[5] `https://pub.uni-bielefeld.de/data/2913177`

[6] `https://opendatacommons.org/licenses/pddl/1.0/`

6. Conclusion

We presented a corpus of multimodal descriptions, in which speech and gestures were used to describe spatial configurations of objects. We described the task designs, recording setups as well as the data annotation scheme. To investigate the usability of the corpus, we also provided preliminary analysis results concerning language, gesture behaviours and multimodal behaviours, then discussed possible use cases of the corpus such as modelling the interpretation of multimodal descriptions and generating multimodal behaviours.

7. Acknowledgments

This work was supported by the China Scholarship Council and the Cluster of Excellence Cognitive Interaction Technology CITEC (EXC 277) at Bielefeld University funded by the German Research Foundation (DFG). This work was also supported by the German Academic Exchange Service (DAAD) and sponsored by the German Federal Ministry of Education and Research (BMBF).

8. Bibliographical References

Alibali, M. (2005). Gesture in spatial cognition: Expressing, communicating, and thinking about spatial information. *Spatial Cognition and Computation*, 5(4):307–331.

Cassell, J., Kopp, S., Tepper, P., Ferriman, K., and Striegnitz, K. (2007). Trading spaces: How humans and humanoids use speech and gesture to give directions. *Conversational informatics*, pages 133–160.

Cheng, H., Yang, L., and Liu, Z. (2016). Survey on 3d hand gesture recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(9):1659–1673.

Emmorey, K., Tversky, B., and Taylor, H. a. (2000). Using space to describe space: Perspective in speech, sign, and gesture. *Spatial Cognition and Computation*, 2(3):157–180.

Fotinea, S.-E., Efthimiou, E., Koutsombogera, M., Dimou, A.-L., Goulas, T., and Vasilaki, K. (2016). Multimodal resources for human-robot communication modelling. In *LREC*.

Han, T., Kennington, C., and Schlangen, D. (2018). Placing Objects in Gesture Space: Toward Real-Time Understanding of Spatial Descriptions. In *Proceedings of the thirty-second AAAI conference on artificial intelligence (AAAI18)*. The association for the advancement of artificial intelligence.

Kelley, J. F. (1983). An empirical methodology for writing user-friendly natural language computer applications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '83, pages 193–196, New York, NY, USA. ACM.

Kendon, A. (1980). Gesticulation and speech: two aspects of the process of utterance. *The Relationship of Verbal and Nonverbal Communication*, 25:207–227.

Kousidis, S., Pfeiffer, T., and Schlangen, D. (2013). MINT . tools : Tools and Adaptors Supporting Acquisition , Annotation and Analysis of Multimodal Corpora. In *Proceedings of Interspeech 2013*, pages 2649–2653, Lyon, France. ISCA.

Liu, L. and Shao, L. (2013). Learning discriminative representations from rgb-d video data. In *IJCAI*, volume 4, page 8.

Lücking, A., Bergmann, K., Hahn, F., Kopp, S., and Rieser, H. (2010). The bielefeld speech and gesture alignment corpus (saga). In *LREC 2010 workshop: Multimodal corpora–advances in capturing, coding and analyzing multimodality*.

Marin, G., Dominio, F., and Zanuttigh, P. (2014). Hand gesture recognition with leap motion and kinect devices. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 1565–1569. IEEE.

Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X.-F., Kirbas, C., McCullough, K. E., and Ansari, R. (2002). Multimodal human discourse: gesture and speech. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 9(3):171–193.

Ragsdale, J. D. and Fry Silvia, C. (1982). Distribution of kinesic hesitation phenomena in spontaneous speech. *Language and Speech*, 25(2):185–190.

Sadeghipour, A. and Morency, L.-P. (2011). 3D Iconic Gesture Dataset.

Schiel, F., Steininger, S., and Türk, U. (2002). The smartkom multimodal corpus at bas. In *LREC*.

Schneider, L. F. and Taylor, H. a. (1999). How do you get there from here? Mental representations of route descriptions. *Applied Cognitive Psychology*, 13(September 1998):415–441.

Tompson, J., Stein, M., Lecun, Y., and Perlin, K. (2014). Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics*, 33, August.