Natural Language Informs the Interpretation of Iconic Gestures: A Computational Approach

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

Abstract

When giving descriptions, speakers often signify object shape or size with hand gestures. Such so-called 'iconic' gestures represent their meaning through their relevance to referents in the verbal content, rather than having a conventional form. The gesture form on its own is often ambiguous, and the aspect of the referent that it highlights is constrained by what the language makes salient. We show how the verbal content guides gesture interpretation through a computational model that frames the task as a multi-label classification task that maps multimodal utterances to semantic categories, using annotated human-human data.

1 Introduction

Besides natural language, human communication often involves other modalities such as hand gestures. As shown in Figure 1, when describing *two lanterns*, one can describe "two lanterns" verbally, while showing the **relative position** with two hands facing each other. Interestingly, when the same gesture is accompanied by the utterance "a ball", the same gesture may indicate **shape**. These gestures (referred to as 'iconic gestures' in gesture studies (McNeill, 1992)) are characterised as conveying meanings through similarity to referents in verbal content, rather than conventional forms of shape/trajectory. Hence, the interpretation of iconic gestures largely depends on verbal content.

Although this theory has been proposed and confirmed in various gesture studies (Feyereisen and De Lannoy, 1991; McNeill, 1992; Kita and Özyürek, 2003; Kita et al., 2007; Özyürek et al., 2008; Bergmann et al., 2014, 2013b), it has not attracted much attention from works on human-computer interfaces (HCIs), which usually assume that gestures have predefined meanings either through conventional agreements (e.g., "thumb up" for "great"), or defined by the system (e.g., "circling" for "circle") (Stiefelhagen et al., 2004; Burger et al., 2012; Lucignano et al., 2013; Rodomagoulakis et al., 2016). Hence, the systems can only interpret a limited number of gestures by classifying gestures based on the shape/trajectory of hands, then combining the information with language. We propose that, in order to incorporate iconic gestures in HCIs, natural language should be taken as an important resource to interpret iconic gestures.

Figure 1: Speech / gesture description of a virtual scene: "*. . . sind halt zwei Laternen*" ("[there] are two lanterns"). Gestures indicate the **amount** (two) and **relative placement** of the two lanterns, while speech indicates the **entity** name and **amount**. From (Lücking et al., 2010).

The relation between speech and iconic gestures has certainly been investigated in previous work. Empirical studies such as (Kita and Özyürek, 2003; Kita et al., 2007) analysed speech and gesture semantics with statistical methods and show that the semantics of speech and gestures coordinate with each other. However, it remains unclear how to computationally derive the semantics of iconic ges-

Verbal utterance U	"two, lanterns"
Gesture G	*two hands facing each other*
Speech semantics	*[entity, amount]*
Gesture semantics	*[relative position, amount]*
Multi-modal semantics	*[entity, relative position, amount]*

Figure 2: Example of a multimodal utterance, and semantic categories.

tures and build corresponding multimodal semantics together with the accompanying verbal content. In this paper, we address this "how" question and present a computational approach that predicts speech and gesture semantic categories using speech and gesture input as features. Speech and gesture information within the same semantic category can then be fused to form a complete multimodal meaning, where previous methods on representing multimodal semantic (Bergmann and Kopp, 2008; Bergmann et al., 2013a; Lascarides and Stone, 2009; Giorgolo, 2010) can be applied. Consequently, this enables HCIs to construct and represent multimodal semantics of natural communications involving iconic gestures.

We investigated whether language informs the interpretation of iconic gestures with the data from the SAGA corpus (Lücking et al., 2010). From the SAGA corpus, we take gesture-speech ensembles as well as semantic category annotations of speech and gestures according to the information they convey. Using words and annotations of gestures to represent verbal content and gesture information, we conducted experiments to map language and gesture inputs to semantic categories. The results show that language is more informative than gestures in terms of predicting iconic gesture semantics and multimodal semantics.

2 Task formulation

We now describe the task formally. Suppose a verbal utterance U is accompanied by a gesture G (as shown in Figure 2), we represent the speech-gesture ensemble as (U, G). The ultimate goal is to map the input information of (U, G) to a set of semantic categories according to the information they convey (as shown in Figure 3), then compose the multi-modal semantics of the ensemble with information in the same category across speech and gestures.

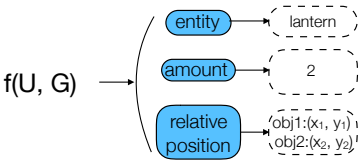

Figure 3: Mapping a speech-gesture ensemble to semantic categories in blue rectangles (U and G indicate speech and gesture). Dashed rectangles indicate the value of each semantic category, which are not included in our current work.

We define a mapping function f that takes a speech-gesture ensemble (U, G) as input, and outputs semantic categories c_i, computed by the set of features of U and G. Additionally, we assume each modality has its own meaning function $f_u(U)$ and $f_g(G)$. In this paper, we make the assumption that *multi-modal meaning* outputted by $f(U, G)$ is in fact the union of $f_u(U)$ and $f_g(G)$:

$$\begin{aligned} f_u(U) &= \{c_1, c_2\} \\ f_g(G) &= \{c_2, c_3\} \\ f(U, G) &= \{c_1, c_2, c_3\} \end{aligned} \quad (1)$$

Figure 3 shows an example of mapping the verbal utterance "two lanterns" to semantic categories {*amount*, *entity*}, while mapping the gesture to categories: {*amount*, *relative position*}. The semantics of the ensemble (U, G) is composed of the semantic categories and their values (in the dashed boxes). In this work we focus on predicting the semantic category rather than their value, which we leave for future work.

We derive input features for the mapping task from speech and gestures respectively:

a) Language features: The word tokens of each verbal utterance are taken as a bag-of-words to represent linguistic information. **b)** Gesture features: Hand movements and forms, including hand shape, palm direction, path of palm direction, palm movement direction, wrist distance, wrist position, path of wrist, wrist movement direction, back of hand direction and back of hand direction movement, are derived as gesture features (as there was no hand motion data, these features were manually annotated, see below for details).

Modelling the learning task We frame the verbal utterance/gesture multimodal semantic category mapping problem as a multi-label classification task (Tsoumakas and Katakis, 2006) where several labels are predicted for an input.

Given an input feature vector $\boldsymbol{X}$, we predict a set of semantic category labels $\{c_1, \cdots, c_i\}$, of which the length is variable. The prediction task can be further framed as multiple binary classification tasks. Technically, we trained a linear support vector classifier (SVC)[1] for each semantic label c_i (6 label classifiers in total). Given an input feature $\boldsymbol{X}$, we apply all semantic label classifiers to the feature vector. If a semantic label classifier gives positive prediction for input $\boldsymbol{X}$, we assign the semantic label to the input. For example, given feature vector of the input utterance "two lanterns", only the *amount* and *entiry* label classifiers give positive predictions, thus we assign *amount* and *entity* to the input utterance.

The word/gesture utterances are encoded as several-hot feature vectors as input of the classifiers, which will be explained now.

3 The SAGA corpus

We conducted the experiments with the SAGA corpus (Lücking et al., 2010), which provides fine-grained annotations for speech and gestures.

The data The corpus consists of 25 dialogues of route and sight descriptions of a virtual town. In each dialogue, a route giver gave descriptions (e.g., route directions, shape, size and location of buildings) of the virtual town to a naive route follower with speech (in German) and gestures. The dialogues were recorded with three synchronised cameras from different perspectives.

In total, 280 minutes of video and audio data were recorded. The audio was manually transcribed and aligned with videos; the gestures were manually annotated and segmented according to video and audio recordings. We selected 939 speech-gesture ensembles out of 973 annotations (Bergmann et al., 2011), omitting 34 without full annotations of speech/gesture semantic categories and gesture features. The semantic categories were annotated according to the semantic information that speech and gestures contained. In our data set, each item is a tuple of 4 elements: *(words, gesture features, speech semantic categories, gesture semantic categories)*.

[1] penalty: $\ell 2$, penalty parameter C=1.0, maximum iteration 1000, using an implementation in `http://scikit-learn.org`.

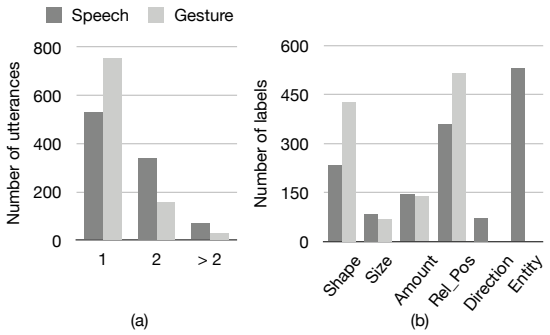

Figure 4: (a) Histogram of semantic labels per utterance/gesture. (b) Histogram of semantic labels. (Rel_Pos indicates relative position.)

There are 5 gesture semantic category labels: *shape*, *size*, *direction*, *relative position*, *amount*; the speech semantic labels consist of these and an extra label of *entity* (6 labels in total). Since there was only one gesture labeled as *direction*, we treat it as a rare instance, and removed it from the evaluation experiments. From these the multi-modal category labels are derived as the union of those two sets for each ensemble.

Data statistics Bergmann et al. (2011) provides detailed data statistics regarding the relation of speech and gestures of the corpus. As we focus on speech and gesture semantics only here, we report statistics only for the 939 speech-gesture ensembles. On average, each verbal utterance is composed of 3.15 words. 386 gestures (41%) provide a semantic category on top of the verbal utterance (e.g., speech: {*amount*, *shape*}, gesture: {*relative position*}), 312 (33%) gestures convey the same amount of semantic information as the verbal utterance (e.g., speech: {*amount*, *shape*}, gesture: {*amount*, *shape*}), and 241 (26%) conveys part of the semantics of the verbal utterance (e.g., speech: {*amount*, *shape*}, gesture: {*amount*}).

As shown in Table 4 (a), 56% of verbal utterances and 80% of gestures are annotated with only a single label. On average, each gesture was annotated with 1.23 semantic labels and each utterance with 1.51 semantic labels. As shown in Figure 4 (b), there are many more utterances labeled with *shape*, *relative*

position and *entity* than the other labels, making the data unbalanced. Moreover, there are considerably more gestures annotated with labels of *shape* and *relative position*.

Gesture features Since there is no tracked hand motion data, we used the manual annotations to represent gestures. For instance, the gesture in Figure 1 is annotated as: Left hand: [5_bent, PAB/PTR, BAB/BUP, C-LW, D-CE]; right hand: [C_small, PTL, BAB/BUP, LINE, MD, SMALL, C-LW, D-CE] in the order of hand shape, hand palm direction, back of hand direction, wrist position. (See (Lücking et al., 2010) for the details of the annotation scheme). Other features such as path of palm direction which are not related to this static gesture were set as 0.

We treated these annotated tokens as "words" that describe gestures. Annotations with more than 1 token were split into a sequence of tokens (e.g., BAB/BUP to BAB, BUP). Therefore, gesture feature sequences have variable lengths, in the same sense as utterances have variable amount of word tokens.

4 Experiments

We randomly selected 70% of the gesture-speech ensembles as a training set, using the rest as a test set. We designed 3 experiments to investigate whether and to what degree language and gestures inform mono-modal and multimodal semantics. Each experiment was conducted under 3 different setups, namely, using: a) only gesture features; b) only language features; c) gesture features and language features, as shown in Table 1.

Metrics We calculated **F1-score**, **precision** and **recall** for each label, and find their average, weighted by the number of true instances for each label, so that imbalanced labels are taken into account.

4.1 Results

Language semantics As shown in Table 1, the most informative features of language semantic categories are words on their own. It achieves an F1-score of 0.79 for each label, well above a chance level baseline accuracy 0.17. While as expected, gesture features are not very informative for language semantics, the gesture-only still classifier outperforms the chance level baseline with 0.38. The combination of features in the joint classifier results in slightly worse performance than language features alone, suggesting some of the gestural semantics may be complementary to, rather than identical to, the language semantics.

Semantics	Features	Precision	Recall	F1-score
Language	L	0.85	0.75	**0.79**
	G	0.47	0.37	0.38
	L+G	0.86	0.69	0.75
Gesture	L	0.80	0.78	**0.78**
	G	0.59	0.63	0.61
	L+G	0.82	0.77	**0.78**
Multimodal	L	0.82	0.80	**0.81**
	G	0.62	0.60	0.58
	L+G	0.83	0.80	0.80

Table 1: Evaluation results. (L and G indicates language and gesture.)

Gesture semantics While language features help predict the semantics of their own modality, the same is not true of gesture features. The language-only classifier achieves an F1-score of 0.78 when predicting gesture semantics, while the gesture features-only setting only achieves 0.61. Combining language and gesture features does not improve performance, but results in a slightly higher precision score (+0.02). This is consistent with previous observations in gesture studies (Feyereisen and De Lannoy, 1991) that iconic gestures are difficult to interpret without speech. Even humans perform poorly on such a task without verbal content.

In our setup, the abstract gesture features might be one of the reasons for poor performance. Only 10 manually annotated categories were used to represent gestures, so these features might not be optimal for a computational model. It is possible that with more accurate gesture features (e.g. motion features), gestures can be better represented and more informative for interpreting gesture semantics.

Multimodal semantics As gestures can add meaningful semantic information not present in concurrent speech, we trained and evaluated classifiers on multimodal semantic categories. We as-

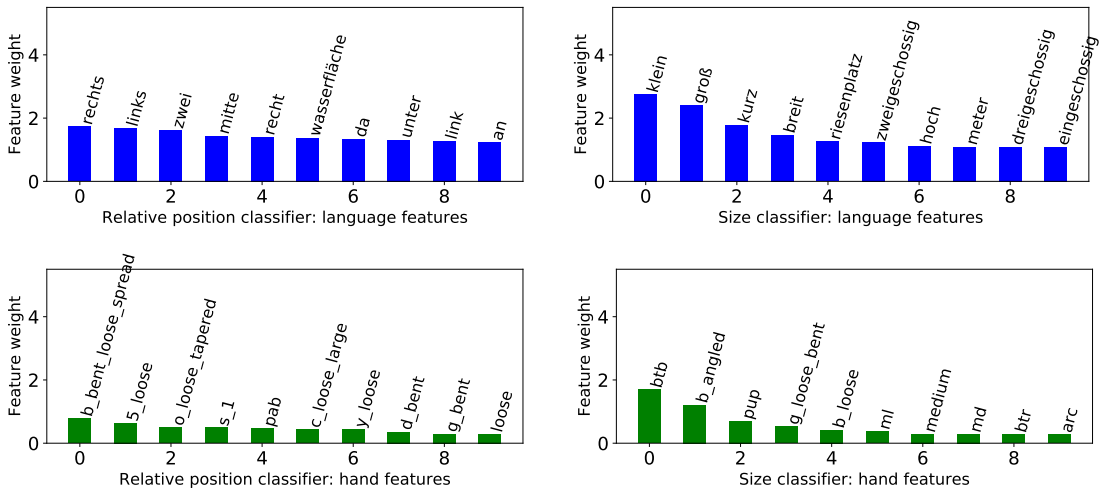

Figure 5: Featuring ranking according to coefficient values (weights assigned to the features).

sume these are the union of the gesture and language semantics for a given ensemble (as in function f in (1) above). As per the data statistics, there are the same possible 6 atomic categories as the language semantics (though they can come from the gesture as well as from the speech). As shown in Table 1, the language-only classifier performs best on this set with an F1-score of 0.81, marginally outperforming the combined language and gesture features system's 0.80. Both significantly outperform the gesture-only classifier. As with the results on gesture semantics, this suggests that multimodal meaning and meaning of iconic gesture relies heavily on speech, in accordance with the finding that the majority of gestures are inherently underspecified semantically by their physical form alone (Rieser, 2015).

Regarding individual semantic categories, we find gesture features are more informative for *shape* and *relative positions*; language is more informative for *size*, *direction* and *amount* in our dataset. Figure 5 shows the gesture and language feature ranking results for classifiers of *entity* and *relative position* accordingly. For *relative position* label prediction, the most informative language features are the words "rechts" (right) and "links" (left), while hand shape (e.g., b_bent_loose_spread, 5_loose) is the most informative gesture feature. For *size* label prediction, the most informative language features are words that specify size such as "klein" (small) and "groß" (big); the most informative gesture features are back of hand palm direction (btb) and hand shape (b_angled).

5 Conclusion

Language and co-verbal gestures are widely accepted as an integral process of natural communication. In this paper, we have shown that natural language is informative for the interpretation of a particular kind of gesture, iconic gestures. With the task of mapping speech and gesture information to semantic categories, we show that language is more informative than gesture for interpreting not only gesture meaning, but also the overall multimodal meaning of speech and gesture. This work is a step towards HCIs which take language as an important resource for interpreting iconic gestures in more natural multimodal communication. In future work, we will predict speech/gesture semantics using raw hand motion features and investigate prediction performance in an online, continuous fashion. This forms part of our ongoing investigation into the interplay of speech and gesture semantics.

Acknowledgements

We are grateful to Kirsten Bergmann and Stefan Kopp for sharing the SAGA corpus. The first author is supported by the China Scholarship Council (CSC). This work was also supported by the Cluster of Excellence Cognitive Interaction Technology CITEC (EXC 277) at Bielefeld University, funded by the German Research Foundation (DFG).

References

Kirsten Bergmann, Volkan Aksu, and Stefan Kopp. 2011. The relation of speech and gestures: temporal synchrony follows semantic synchrony. In *Proceedings of the 2nd Workshop on Gesture and Speech in Interaction (GeSpIn 2011)*.

Kirsten Bergmann, Florian Hahn, Stefan Kopp, Hannes Rieser, and Insa Röpke. 2013a. Integrating gesture meaning and verbal meaning for german verbs of motion: Theory and simulation. In *Proceedings of the Tilburg Gesture Research Meeting (TiGeR 2013)*.

Kirsten Bergmann, Sebastian Kahl, and Stefan Kopp. 2013b. Modeling the semantic coordination of speech and gesture under cognitive and linguistic constraints. In *International Workshop on Intelligent Virtual Agents*. Springer, pages 203–216.

Kirsten Bergmann, Sebastian Kahl, and Stefan Kopp. 2014. How is information distributed across speech and gesture? a cognitive modeling approach. *Cognitive Processing, Special Issue: Proceedings of KogWis* pages S84–S87.

Kirsten Bergmann and Stefan Kopp. 2008. Multimodal content representation for speech and gesture production. In *Proceedings of the 2nd Workshop on Multimodal Output Generation*. pages 61–68.

B. Burger, I. Ferrané, F. Lerasle, and G. Infantes. 2012. Two-handed gesture recognition and fusion with speech to command a robot. *Autonomous Robots* 32(2):129–147.

Pierre Feyereisen and Jacques-Dominique De Lannoy. 1991. *Gestures and speech: Psychological investigations*. Cambridge University Press.

Gianluca Giorgolo. 2010. *Space and Time in Our Hands*. Ph.D. thesis, Netherlands Graduate School of Linguistics.

Sotaro Kita and Asli Özyürek. 2003. What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and language* 48(1):16–32.

Sotaro Kita, Asli Özyürek, Shanley Allen, Amanda Brown, Reyhan Furman, and Tomoko Ishizuka. 2007. Relations between syntactic encoding and co-speech gestures: Implications for a model of speech and gesture production. *Language and cognitive processes* 22(8):1212–1236.

Alex Lascarides and Matthew Stone. 2009. A formal semantic analysis of gesture. *Journal of Semantics* 26(4):393–449.

Lorenzo Lucignano, Francesco Cutugno, Silvia Rossi, and Alberto Finzi. 2013. A dialogue system for multimodal human-robot interaction. In *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, pages 197–204.

Andy Lücking, Kirsten Bergmann, Florian Hahn, Stefan Kopp, and Hannes Rieser. 2010. The bielefeld speech and gesture alignment corpus (saga). In *LREC 2010 workshop: Multimodal corpora–advances in capturing, coding and analyzing multimodality*.

David McNeill. 1992. Hand and Mind: What Gestures Reveal About Thought .

Asli Özyürek, Sotaro Kita, Shanley Allen, Amanda Brown, Reyhan Furman, and Tomoko Ishizuka. 2008. Development of cross-linguistic variation in speech and gesture: Motion events in english and turkish. *Developmental psychology* 44(4):1040.

Hannes Rieser. 2015. When hands talk to mouth. gesture and speech as autonomous communicating processes. *SEMDIAL 2015 goDIAL* page 122.

Isidoros Rodomagoulakis, Nikolaos Kardaris, Vassilis Pitsikalis, E Mavroudi, Athanasios Katsamanis, Antigoni Tsiami, and Petros Maragos. 2016. Multimodal human action recognition in assistive human-robot interaction. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, pages 2702–2706.

R. Stiefelhagen, C. Fugen, R. Gieselmann, H. Holzapfel, K. Nickel, and A. Waibel. 2004. Natural human-robot interaction using speech, head pose and gestures. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566)*. volume 3, pages 2422–2427.

Grigorios Tsoumakas and Ioannis Katakis. 2006. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining* 3(3).