

Learning to Interpret and Apply Multimodal Descriptions

Ting Han

Doctor of Philosophy
Department of Linguistics and Literature
Bielefeld University

2018

Copyright ©2018 Ting Han, Bielefeld, Germany.

Dissertation zur Erlangung des akademischen Grades *Doctor philosophiae* (Dr. phil.) vorgelegt an der Fakultät für Linguistik und Literaturwissenschaft der Universität Bielefeld am 25 Oktober 1982.

Prüfungskommission:

Prof. Dr. David Schlangen (Betreuer und Gutachter)

Prof. Dr. Petra Wagner (Gutachter)

Prof. Dr. Stefan Kopp

Datum der mündlichen Prüfung: 12 Juni 2018.

Typeset mit L^AT_EX.

⊗ Printed on acid-free, aging-resistant paper (according ISO 9706). Gedruckt auf säure-freiem, alterungsbeständigen Papier (nach ISO 9706).

Abstract

Enabling computers to understand natural human communication is a goal researchers have been long aspired to in artificial intelligence. Since the concept demonstration of “*Put-That-There*” in 1980s, significant achievements have been made in developing multimodal interfaces that can process human communication such as speech, eye gaze, facial emotion, co-verbal hand gestures and pen input. State-of-the-art multimodal interfaces are able to process pointing gestures, symbolic gestures with conventional meanings, as well as gesture commands with pre-defined meanings (e.g., *circling* for “select”). However, in natural communication, co-verbal gestures/pen input rarely convey meanings via conventions or pre-defined rules, but embody meanings relatable to the accompanied speech.

For example, in route given tasks, people often describe landmarks verbally (e.g., *two buildings*), while demonstrating the **relative position** with two hands facing each other in the space. Interestingly, when the same gesture is accompanied by the utterance *a ball*, it may indicate the **size** of the ball. Hence, the interpretation of such co-verbal hand gestures largely depends on the accompanied verbal content. Similarly, when describing objects, while verbal utterances are most convenient for conveying meanings symbolically (e.g., describing **colour** and **category** with the utterance “*a brown elephant*”), hand-drawn sketches are often deployed to convey **iconic** information such as the exact shape of the elephant’s trunk, which is typically difficult to encode in language.

This dissertation concerns the task of learning to interpret multimodal descriptions composed of verbal utterances and hand gestures/sketches, and apply corresponding interpretations to tasks such as image retrieval. Specifically, I aim to address following research questions: **1)** For co-verbal gestures that embody meanings relatable to accompanied verbal content, how can we use natural language information to interpret the semantics of such co-verbal gestures, e.g., does a gesture indicate relative position or size? **2)** As an integral system of communication, speech and gestures not only bear close semantic relations, but also close temporal relations. To what degree and on which dimensions can hand gestures benefit the interpretation of multimodal descriptions? **3)** While it’s obvious that iconic information in hand-drawn sketches enriches the verbal content in object descriptions, how to model the joint contributions of such multimodal descriptions and to what degree can verbal descriptions compensate reduced iconic details in hand-drawn sketches?

To address the above research questions, in this dissertation, I first introduce three multimodal description corpora: a spatial description corpus composed of natural language and placing gestures (also referred as abstract deictics), a multimodal object description corpus composed of natural language and hand-drawn sketches, and an existing corpus - the Bielefeld

Speech and Gesture Alignment Corpus (SAGA), which provides fine-grained annotations of speech and hand gestures in a route giving and following task.

After introducing the corpora related to studies in this dissertation, I first describe a system that models the interpretation and application of spatial descriptions and explored three variants of representation methods of the verbal content. When representing the verbal content in the descriptions with a set of automatically learned symbols, the system's performance is on par with representations with manually defined symbols (e.g., pre-defined object properties), showing that besides learning to interpret and apply multimodal spatial descriptions, the system can also learn to automatically represent the multimodal content. Moreover, I show that abstract deictic gestures not only lead to better understanding of spatial descriptions, but also result in earlier correct decisions of the system, which can be used to trigger immediate reactions in dialogue systems.

Going beyond deictics in multimodal descriptions, I also investigated the interplay of semantics between symbolic (natural language) and iconic (sketches) modes in multimodal object descriptions, where natural language and sketches jointly contribute to the communications. I model the meaning of natural language and sketches two existing models and combine the meanings from both modalities with a late fusion approach. The results show that even adding reduced sketches (30% of full sketches) can help in the retrieval task. Moreover, in current setup, natural language descriptions can compensate around 30% of reduced sketches.

In the above tasks, I modelled the interpretation of multimodal descriptions composed of deictic and iconic elements separately. Deictic and iconic elements were represented with different methods, assuming that the system automatically knows how to represent deictic and iconic content (i.e., extracting position information from deictics while encoding drawing trajectories as vectors). In a more realistic setup, a system should learn to resolve how to represent the semantics of hand gestures. I frame the problem of learning gesture semantics as a multi-label classification task using natural language information and hand gesture features. I describe an experiment conducted with the SAGA corpus, and show that natural language is informative for the learning the semantics of verbal utterances and hand gestures.

Acknowledgements

Without the support from many people, I could not have finished writing this dissertation.

First of all, I would like to thank my *Doktorvater* Prof. David Schlangen for his continuous help, support and encouragement throughout my PhD study. In the past 4 years, he was always available for discussions and giving feedback. During our meetings, he always gives insightful advices and inspires me with curiosity. I am really grateful that he was so patient to listen to my naive thoughts, guide me, help with my incoherent papers and presentations, and so on. He is definitely one of the best advisers I have ever met, and also an excellent role model for me.

I am grateful to China Scholarship Council for providing me with a 4-year scholarship. Without the funding, I could not do the research that interests me. I would also like to thank the Cluster of Excellence Cognitive Interaction Technology (CITEC) at Bielefeld University for providing me with a travel funding, which enables me to attend conferences. I'd like to thank the Rectorate of Bielefeld University for providing me with a bridge funding from the Bielefeld Young Researchers' Fund which supports me to finish the dissertation writing after my PhD scholarship ended.

I am also grateful to my colleagues in the Dialogue Systems Group (in no particular order): Soledad Lopez, Dr. Sina Zariß, Dr. Julian Hough, Dr. Casey Kennington, Dr. Iwan de Kok, Dr. Spyros Kousidis, and Nikolai Llinikh. They have kept me motivated throughout my PhD, and always give insightful feedback on my projects and presentations. Dr. Julian Hough and Dr. Sina Zariß are especially among the best co-authors one could work with. They both contributed enormously to our joint projects, and demonstrated to me how to turn ideas to research projects.

Thanks should also go to our student assistants: Kai Mismahl, Michael Bartholdt, Oliver Eickmeyer, and Gerdis Anderson, for helping in the data collection and data annotations. Kai and Michael picked me up at the train station on the morning of a rainy day and helped me to settle down in Bielefeld. As a person who always gets lost, I probably couldn't manage it without their help.

Enormous thanks to my friends and family for their support during the journey of my PhD. This thesis would not have been possible without their encouragement and invaluable help.

Relevant Publications

Parts of this thesis have appeared previously in the following publications:

- **Ting Han**, Casey Kennington, and David Schlangen. Placing objects in gesture space: Towards real-time understanding of multimodal spatial descriptions. *In Thirty-second AAAI conference on artificial intelligence (AAAI18)*, 2018.
- **Ting Han** and David Schlangen. A corpus of natural multimodal spatial scene descriptions. *In 11th edition of the Language Resources and Evaluation Conference (LREC18)*, 2018.
- **Ting Han**, Julian Hough, and David Schlangen. Natural language informs the interpretation of iconic gestures: a computational approach. *In The 8th International Joint Conference on Natural Language Processing (IJCNLP17)*, 2017.
- **Ting Han** and David Schlangen. Draw and tell: Multimodal descriptions outperform verbal- or sketch-only descriptions in an image retrieval task. *In The 8th International Joint Conference on Natural Language Processing (IJCNLP17)*, 2017.
- **Ting Han**, Casey Kennington, and David Schlangen. Building and Applying Perceptually-Grounded Representations of Multimodal Scene Descriptions. *In Proceedings of the 19th SemDial Workshop on the Semantics and Pragmatics of Dialogue (goDIAL)*, 2015.
- **Ting Han**, Spyridon Kousidis, and David Schlangen. A corpus of virtual pointing gestures. *In The RefNet Workshop on Psychological and Computational Models of Reference Comprehension and Production*, 2014. [poster]
- **Ting Han**, Spyridon Kousidis, and David Schlangen. Towards automatic understanding of ‘virtual pointing’ in interaction. *In Proceedings of the 18th SemDial Workshop on the Semantics and Pragmatics of Dialogue (DialWatt), Posters, pages 188–190*, 2014. [poster]

Declaration

I declare that I am aware of the doctoral degree regulations as specified by the Rahmenpromotionsordnung of Bielefeld University (published June 15 2010) as well as the regulations as specified by the Promotionsordnung of the Faculty of Linguistics and Literary Studies of Bielefeld University (published April 1 2015). I further declare that this thesis was written by myself and that the work contained therein is my own, except in those cases where it is explicitly stated otherwise in the text. No payment or payment-in-kind was made to third parties in any form for any work related to the graduation proceedings. This work has only been submitted to Bielefeld University and has not been submitted to another degree, another scientific examination, or another university as a dissertation.

(Ting Han)

Kurzfassung

Schon lange strebt die Forschung in der Künstlichen Intelligenz danach, Maschinen zu entwickeln, die natürliche menschliche Kommunikation verstehen. Seit der ersten Entwicklung der Idee von "Put That There" in den 1980ern wurden signifikante Fortschritte bei der Implementierung von multimodalen Schnittstellen gemacht, die menschliche Kommunikation via gesprochener Sprache, Blickbewegungen, Mimik, ko-verbalen Gesten und Zeichnungen mit Stift verarbeiten. Auf dem aktuellen Stand der Forschung sind multimodale Schnittstellen in der Lage, gesprochene Sprache, deiktische Gesten, symbolische Gesten mit konventionalisierter Bedeutung zu verarbeiten, sowie gestenbasierte Kommandos mit vordefinierten Bedeutungen (z.B. Kreisen für "Auswahl"). Allerdings verwenden Sprecher in natürlicher Kommunikation oft Gestik oder Zeichnungen, die nicht konventionalisiert sind und keine vordefinierte Bedeutung tragen, sondern die nur relativ zur gleichzeitig gesprochenen Sprache Bedeutung zum Ausdruck bringen.

Ein Beispiel sind Wegerklärungsaufgaben, bei denen Sprecher oft Orientierungspunkte beschreiben (zwei Kirchen) und ihre relative Position mit zueinander ausgerichteten Händen demonstrieren. Interessanterweise kann die gleiche Geste, wenn sie mit einer Äußerung wie der Ball auftritt, auf die Größe des Balls hindeuten. Daher ist die Interpretation von ko-verbalen Gesten zu einem großen Teil abhängig vom verbalen Inhalt der gleichzeitig kommuniziert wird. In ähnlicher Weise können bei der Beschreibung von Objekten (ein brauner Elefant) zum Beispiel handgezeichnete Skizzen verwendet werden um ikonische Information wie die exakte Form des Elefantenüssels zu übermitteln, während natürliche Sprache in diesem Fall eher geeignet ist, Information wie Farbe oder Kategorie des Objekts zu beschreiben.

Diese Dissertation beschäftigt sich damit, multimodale Äußerungen, die aus nat"urlicher Sprache und ko-verbalen Gesten bestehen, zu interpretieren und in Anwendungen wie z.B. image retrieval auszunutzen. Wir bearbeiten dabei folgende Forschungsfragen: 1) Wie können wir natürliche Sprache nutzen um die Semantik ko-verbaler Gesten vorherzusagen, da doch die Interpretation ko-verbaler Gesten vom verbalen Inhalt abhängt? 2) Als integrale Bestandteile von Kommunikation stehen Sprache und Gestik nicht nur in enger semantischer Beziehung, sondern auch in temporaler. In welchem Maße und in welchen Dimensionen können Gesten dabei helfen, multimodale Beschreibungen zu interpretieren? 3) Während es einerseits offensichtlich ist, dass ikonische Information in handgezeichneten Skizzen den verbalen Inhalt von Objektbeschreibungen anreichert, stellt sich andererseits die Frage, wie diese multimodalen Beschreibungen modelliert werden und in welchem Maße verbale Beschreibungen reduzierte ikonische Information kompensieren können. Um diese Fragen zu untersuchen, führen wir zunächst drei multimodale Korpora mit Objektbeschreibungen ein: räumliche Beschreibungen

gen bestehend aus natürlicher Sprache und platzierenden Gesten (auch als abstrakte deiktische Gesten bezeichnet), multimodale Beschreibungen bestehend aus natürlicher Sprache und handgezeichneten Skizzen und ein existierendes Korpus - den Bielefeld Speech and Gesture Alignment Corpus (SAGA) (Kapitel 3). Dann operationalisieren wir das Problem des Lernens von Gestensemantik mittels *Multi-Label*-Klassifikation basierend auf natürlicher Sprache und annotierten Merkmalen der Handgesten und zeigen das natürliche Sprache die Interpretation von Handgesten informiert.

Außerdem beschreiben wir ein System, das die Interpretation und die Anwendungen von räumlichen Beschreibungen modelliert und explorieren dabei 3 Varianten von Repräsentationen des verbalen Inhalts. Wir zeigen, dass abstrakte deiktische Gesten nicht nur zu einem besseren Verständnis von räumlichen Beschreibungen beitragen, sondern auch zu früheren korrekten Entscheidungen des Systems führen, was ausgenutzt werden kann, um unmittelbare Reaktionen in Dialogsystemen zu implementieren.

Schließlich untersuchen wir die semantischen Interaktionen von symbolischer (natürlicher-sprachlicher) und ikonischer (auf Zeichnungen basierender) Modalität in multimodalen Objektbeschreibungen, bei denen natürliche Sprache und ikonische Information gemeinsam zur Bedeutung der Beschreibung beitragen. Wir modellieren die Bedeutung natürlicher Sprache und Skizzen mit zwei existierenden Modellen und kombinieren deren Bedeutungsrepräsentationen mit einem *late fusion*-Ansatz. Wir zeigen, dass sogar reduzierte Skizzen positiv zur Performanz des Retrieval-Systems beitragen. Zudem können in diesem Ansatz verbale Beschreibungen bis zu 30% der reduzierten Skizzen kompensieren.

Contents

| | | |
|----------|--|-----------|
| 1 | Introduction | 16 |
| 1.1 | Tasks | 18 |
| 1.2 | Thesis outline | 21 |
| 2 | Related work | 24 |
| 2.1 | Speech and gestures in natural communications | 24 |
| 2.2 | Typologies of hand gestures | 26 |
| 2.2.1 | Abstract deictics in spatial descriptions | 28 |
| 2.2.2 | Describe objects with iconic gestures/sketches | 29 |
| 2.3 | Relations between speech and co-verbal hand gestures | 30 |
| 2.3.1 | Semantic coordinations between co-verbal gestures and verbal content | 30 |
| 2.3.2 | Temporal alignment between gestures and speech | 31 |
| 2.4 | Multimodal human-computer interfaces | 31 |
| 2.4.1 | Natural language processing | 34 |
| 2.4.2 | Gesture recognition and interpretation | 35 |
| 2.4.3 | Multimodal fusion | 37 |
| 2.5 | Representation of multimodal content | 39 |
| 2.6 | Existing multimodal datasets | 41 |
| 2.7 | Summary | 43 |
| 3 | Multimodal corpora | 44 |
| 3.1 | Multimodal spatial scene description corpus | 44 |
| 3.1.1 | The scene description experiment | 46 |
| 3.1.2 | The spatial description experiment | 50 |
| 3.2 | Multimodal object description corpus | 56 |
| 3.2.1 | The Sketchy dataset | 57 |
| 3.2.2 | Augmenting sketches with verbal descriptions | 58 |
| 3.2.3 | Data statistics | 60 |

| | |
|---|-----------|
| <i>CONTENTS</i> | 11 |
| 3.3 The SAGA corpus | 61 |
| 3.4 Summary | 63 |
| 4 A system of understanding multimodal spatial descriptions | 64 |
| 4.1 Modelling the interpretation of multimodal spatial descriptions | 64 |
| 4.2 System overview | 66 |
| 4.2.1 Utterance segmentation | 68 |
| 4.2.2 Representing scene descriptions | 68 |
| 4.2.3 Applying gestural information | 69 |
| 4.3 Learning knowledge from prior experience | 71 |
| 4.3.1 The TAKE corpus | 71 |
| 4.3.2 Learning mappings to logical forms | 72 |
| 4.3.3 Learning perceptual groundings | 73 |
| 4.4 Applying the represented knowledge | 75 |
| 4.5 Experiment | 76 |
| 4.5.1 A scene description corpus | 76 |
| 4.5.2 Evaluation | 77 |
| 4.5.3 Results | 77 |
| 4.6 Summary | 78 |
| 5 Towards real-time understanding of multimodal spatial descriptions | 80 |
| 5.1 Real-time understanding of spatial scene descriptions | 80 |
| 5.2 System overview | 81 |
| 5.2.1 Gesture detection | 83 |
| 5.2.2 Gesture interpretation | 85 |
| 5.2.3 Utterance segmentation | 87 |
| 5.2.4 Natural language understanding | 87 |
| 5.2.5 Multimodal fusion & application | 88 |
| 5.3 System evaluation | 90 |
| 5.3.1 Gesture detector evaluation | 90 |
| 5.3.2 Gesture interpretation evaluation | 90 |
| 5.3.3 Utterance segmentation evaluation | 91 |
| 5.3.4 Whole system evaluation | 91 |
| 5.3.5 Incremental evaluation | 94 |
| 5.3.6 Human understanding | 96 |
| 5.4 Summary | 97 |

| | | |
|----------|---|------------|
| 6 | Investigate symbolic and iconic modes in object descriptions | 98 |
| 6.1 | Draw and Tell: iconic and symbolic modes in object descriptions | 98 |
| 6.2 | Model the meaning of multimodal object descriptions | 100 |
| 6.2.1 | Grounding verbal descriptions | 100 |
| 6.2.2 | Comparing sketches with images | 101 |
| 6.2.3 | Fusion | 102 |
| 6.3 | Experiments | 102 |
| 6.3.1 | The image retrieving task | 102 |
| 6.3.2 | Metrics | 103 |
| 6.3.3 | Experiment 1: Mono-modal models | 103 |
| 6.3.4 | Experiment 2: multimodal models | 104 |
| 6.3.5 | Experiment 3: reduced sketch details | 105 |
| 6.4 | Discussion | 107 |
| 6.5 | Summary | 107 |
| 7 | Learning semantic categories of multimodal descriptions | 109 |
| 7.1 | Represent multimodal utterances with semantic concepts | 109 |
| 7.2 | Task formulation | 110 |
| 7.3 | Modelling the learning of multimodal semantics | 112 |
| 7.4 | Experiments | 113 |
| 7.4.1 | Language semantics | 114 |
| 7.4.2 | Gesture semantics | 114 |
| 7.4.3 | Multimodal semantics | 114 |
| 7.5 | Summary | 116 |
| 8 | Conclusion and future work | 117 |
| 8.1 | Overview of the dissertation | 117 |
| 8.2 | Future work | 119 |
| | References | 121 |

List of Figures

| | | |
|------|--|----|
| 1.1 | A photograph described with a sketch (on the right) and the utterance “ <i>an elephant, trunk coiled towards mouth</i> ”. | 19 |
| 2.1 | Kendon’s Continuum. As one moves from left to right, the degree of semantic relations between speech and gestures decreases, while the degree of a gesture shows the properties of a language increases. | 28 |
| 3.1 | Spatial layout of landmarks in Example (1). | 45 |
| 3.2 | Providing a description in the Scene Description Experiment. | 46 |
| 3.3 | Leap sensor. | 47 |
| 3.4 | Data statistics of the Scene Description experiment. | 51 |
| 3.5 | Providing a multimodal spatial scene description. | 52 |
| 3.7 | Example of a multimodal description: <i>facing right, trunk coiled toward mouth</i> . | 57 |
| 3.8 | Discriminative description of the left-most photograph provided by crowdworker: <i>facing right, trunk coiled toward mouth</i> . | 59 |
| 3.9 | Example of data validation test. | 60 |
| 3.10 | (a) Histogram of semantic labels per utterance/gesture. (b) Histogram of semantic labels. (Rel.Pos indicates relative position.) | 62 |
| 4.1 | Scene example. | 65 |
| 4.2 | Overview of the system framework. (Modules in grey boxes are not implemented in this chapter, but simulated. See the Experiment section for details.) | 67 |
| 4.3 | Processing pipeline | 69 |
| 4.4 | Example of a good mapping (top) and bad mapping (bottom), numbered IDs represent the perceived objects, the letter IDs represent the described objects. | 70 |
| 4.5 | Example TAKE scene used for training. | 71 |
| 4.6 | Clusters of words according to the co-efficient values of corresponding word classifiers. See 4.3.2 for detailed descriptions. | 74 |
| 4.7 | Simplified (and constructed) pipeline example. The description “here a red T” with gesture at point (1, 3) is represented and mapped to the perceived scenes. Each variant assigns a higher probability to the correct scene, represented by X_2 | 75 |
| 5.1 | Overview of the system. | 81 |
| 5.2 | Examples of stroke hold detection. We used <i>palm magnitude</i> to show the stroke hold phase as it is one of the major features which distinguish stroke hold from other hand movements. | 84 |

| | | |
|-----|--|-----|
| 5.3 | Mapping deictics from gesture space to scene coordinate system. (a) deictic gestures in the gesture space (Leap sensor coordinate system); (b) gestures are mapped to the target scene; (c) gestures are mapped to a distractor scene with different spatial configurations. | 85 |
| 5.4 | Illustration of multimodal fusion & application, given a candidate scene C and following description: U_1 : here G_1 is a small red square, U_2 : here G_2 is a yellow circle. (For clarity of descriptions, the numbers are constructed and not actual computations for this input.) | 89 |
| 5.5 | Average MRR of incremental evaluation. | 94 |
| 5.6 | Results of incremental evaluation. See text for description of metrics. For all metrics, lower numbers denote better performance. | 95 |
| 6.1 | A photograph; a verbal description of its content; and a sketch. | 99 |
| 6.2 | The GoogLe network from Sangkloy et al. (2016). The <i>Image network</i> and the <i>sketch network</i> are both pre-trained with an image/sketch classification task (with classification losses), then fine-tuned for an sketch based image retrieval task with an embedding loss. \mathbf{P} and \mathbf{S} indicate the feature vectors that represent images and sketches. For detailed descriptions of the network and the training procedure, please refer to the original paper. | 101 |
| 6.3 | Retrieval with verbal description only (1st column), verbal description plus 30% sketch (2nd column), 30% sketch (3rd column) and 100% sketch (4th column). | 106 |
| 7.1 | Speech / gesture description of a virtual scene: “... <i>sind halt zwei Laternen</i> ” (“[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). | 110 |
| 7.2 | Example of a multimodal utterance, and semantic categories. | 111 |
| 7.3 | 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. | 112 |
| 7.4 | Featuring ranking according to coefficient values (weights assigned to the features, see Lücking et al. (2010) for the details of the annotation scheme). | 115 |

List of Tables

| | | |
|-----|---|-----|
| 2.1 | Overview of multimodal fusion approaches and respective characteristics. | 38 |
| 4.1 | Overview of representation variants A-C. | 68 |
| 4.2 | Results of the Experiments. Exp. 1: objects in same spatial configuration in all scenes (per retrieval task); Exp. 2: objects potentially in different configurations in scenes, but same three objects in all scenes; Exp. 3: potentially different objects and different locations in all scenes. | 78 |
| 5.1 | Evaluation results of utterance segmenter. | 91 |
| 5.2 | Results of whole system evaluation. | 91 |
| 6.1 | Average recall at K=1 and 10, at different levels of sketch detail. Highest number in column in bold. Numbers for language-only conditions do not change with level of sketch detail. | 105 |
| 7.1 | Evaluation results. (L and G indicates language and gesture.) | 113 |

1

Introduction

Human communication is multimodal in nature. While language is convenient and intuitive for conveying symbolic information, other modalities are often involved in situated communications to complement/supplement verbal content. Hands, being readily available to almost everyone, are often deployed for such purposes.

For example, when giving route descriptions, people often describe the landmarks verbally, while placing hands in the shared space to demonstrate the spatial relations of landmarks, such as:

(1) *here* [*deictic*] *is the bus stop*, [*deictic*] *a bit left of it is a restaurant ...*

by placing hands in the space (conventionally referred as *abstract deictic gestures* (McNeill et al., 1993)), a speaker maps spatial layout of the landmarks from his/her mental image to the shared space. Together with the verbal descriptions, a listener can build his/her mental representation of the landmarks, later navigating itself with the represented knowledge.

Besides the spatial layout, a route giver often also provides detailed visual descriptions of the landmarks such as shape and orientation. A convenient way to describe the contour of a building is to use the gesture space as a canvas and roughly draw the shape of the referent in the space (Cassell et al., 2007). For example:

(2) *the cafeteria in a bell-shaped* _[drawing] *building*

while the verbal description specifies the entity name, the trajectory of the drawing gesture (i.e., iconic gesture) visually signifies the shape of the building.

In gesture studies, abstract deictics and iconic gestures are conventionally referred as *representational gestures*. Representational gestures often reflect conceptual demands of a speaker (Hostetter et al., 2007). Together with natural language, they help a speaker to constitute thoughts (Kita, 2000) and facilitate communication by conceptualising underlying mental representations. Such gestures also enhance listeners' comprehension (Kita, 2000; Alibali, 2005; Beattie and Shovelton, 1999) as the joint meaning of a multimodal utterance occurs in an organised manner and distributes across both modalities (Bergmann et al., 2014).

Note that iconic gestures are not the only approach to resemble visual similarity in natural communications. When a pen and a canvas (e.g., a piece of paper or a painting board) are at hand, one could also illustrate the shape of an object with hand-drawn sketches. Similar to iconic gestures, sketches can also supplement verbal utterances to form a mental representation of the described object.

Although sketches are similar to iconic gestures in the sense of conveying iconic information, there are significant differences between them. Due to the abstract nature of hand gestures and timing pressure in situated communications, iconic gestures usually only signify salient parts of objects. Consequently, iconic gestures bear closer temporal and semantic relations with accompanied verbal content. Thus, the meaning of iconic gestures is relatable to the accompanying verbal content. In comparison, as sketches drawn on a real canvas are static, they can encode more details than iconic gestures and are only loosely related to accompanied content on the temporal level.

As an integral part of human communication, hand gestures have motivated various studies across disciplines. Researchers have investigated the temporal and semantic relations between speech and gestures through empirical studies (Kendon, 1997; McNeill, 2005). These works not only shed light on the interplay of speech gestures in natural human behaviours, but also help to form theoretical hypothesis in computationally modelling multimodal behaviours in natural human communication. To computationally construct the meaning of multimodal communications, multimodal semantic models have been proposed to explore the representation of multimodal semantics in computational systems, providing insights of building and applying the interpretations of speech and co-verbal hand gestures (Lascarides and Stone, 2009; Giorgolo, 2010).

While humans can easily understand the above mentioned multimodal communication, represent the content in their mind, and probably later apply the knowledge to perform real-life

tasks, it remains a challenging task for computers to understand such communication as humans do in artificial intelligence. Researchers in the multimodal human-computer interfaces (HCIs) community have made prominent achievements on enabling computers to understand speech and hand gestures, but only limited to a set of gestures with conventional meanings (Karam and Schraefel, 2005; Turk, 2014) or pre-defined gesture commands, rather than natural representational gestures.

This dissertation aims to explore the interpretation and application of multimodal descriptions composed of natural language and representational gestures/hand-drawn sketches as discussed above. More specifically, I investigate how to learn semantic concepts of representation gestures, how (abstract) deictic gestures facilitate better interpretation of spatial descriptions, and how iconic information together with natural language descriptions encode richer information than language alone. This dissertation contributes to building natural multimodal human-computer interfaces that goes beyond understanding symbolic gestures and deictic gestures.

1.1 Tasks

In this dissertation, I intend to model the interpretation and application of multimodal descriptions. Specifically, I focus on multimodal descriptions composed of representational gestures and hand-drawn sketches.

First of all, I model the interpretation and application of multimodal descriptions composed of *deictics* and verbal utterances. To this end, I started with a task of interpreting spatial scene descriptions, in which abstract deictics supplement the verbal content with spatial layout information.

When describing several landmarks that are not in the situated environment, humans often accompany natural language descriptions with deictic gestures, demonstrating the relative positions with hands in the space. For instance, to help a person to locate a hotel not in current view, a route giving description might be:

- (3) “*Here*_[deixis] *is the train station, [deixis]here is the bus stop, and next to it*_[deixis] *is the hotel.*”

While the verbal utterances indicate the *entity name* (e.g., train station) and *relative position* (e.g., next to it), deictic gestures visually indicate the spatial configurations which complement the verbal content. For example, although the phrase “*next to*” indicates relative positions of the landmarks, the spatial layout between the two landmarks is still unclear. As given the description, a listener still cannot figure out whether the bus stop is to the right or left of the hotel. In this case, the deictic gestures complement speech with concrete spatial layout

information, and consequently result in a clearer route description.

The deictic gestures in the above description are referred as *abstract deictics* in gesture studies (McNeill et al., 1993), featured for placing abstract referents in the gesture space and mapping mental spatial configurations to the shared space. Only together with accompanied speech, the deictic gestures can have determined meanings.

To model the interpretation of spatial descriptions, I started with an empirical study of such descriptions with a simplified setup. Participants were asked to describe several geometric objects and their relative spatial configurations. With the collected corpus, I first explored 3 methods of representing the multimodal descriptions in a multimodal system, then modelled the interpretation of multimodal descriptions and applications with a real-time system and evaluated the system on the general and incremental level.



Figure 1.1: A photograph described with a sketch (on the right) and the utterance “*an elephant, trunk coiled towards mouth*”.

Comparing to deictics, sketches enrich the verbal content with shape information, which convey meanings by assemble visual similarities rather than position information. Such iconic information is typically difficult to describe symbolically with verbal descriptions. For instance, as shown in Figure 1.1, it is a bit ambiguous when there are several elephants with trunks coiled, *drawing a trajectory* to show how the trunk is coiled.

Although iconic gestures share similar nature of conveying visual information, modelling such descriptions with computational approaches requires large scale corpora. Due to technical challenges, collecting such a corpus with hand motion data is not feasible currently. Therefore, in this dissertation, I focus on the task of interpreting object descriptions composed of natural language and hand-drawn sketches, leaving it as future work to model the interpretation task of iconic gestures. Comparing to iconic gestures, it’s easier to collect sketches with detailed timing and path information of drawing strokes by saving them as SVG files and rendering into images, making the data amenable for computational models in the computer vision tasks such as deep neural networks.

To investigate how iconic information facilitates natural communications, I first collected a corpus of real-life photograph descriptions from English speakers using the Crowdfower service,¹. The photographs were selected from ImageNet, and paired with hand-drawn sketches from an existing corpus – the Sketchy dataset, originally introduced in Sangkloy et al. (2016). Note that the language descriptions and sketches were collected separately (see details in the data description in Chapter 3), the temporal relations between language and drawing strokes are not available. As the current study focus on the semantics, I leave it as future work to investigate temporally aligned multimodal descriptions.

I investigated the interplay of symbolic and iconic modes in object descriptions, with sketches representing iconicity of objects and natural language representing symbolic information. Mono-modal and multimodal experiments were designed to evaluate the contributions of symbolic and iconic modes with an image retrieving task, which shows multimodal descriptions outperform mono-modal descriptions.

While full sketches are informative as they encode detailed iconic information, in natural communications, humans often only gesture for the most salient part of an object due to timing pressure. Therefore, it's an interesting question that to what degree the reduced sketch details can be covered by natural language? To this end, I designed multimodal experiments with reduced details of sketches and evaluated the image retrieval performance. The results show that around 30% reduced details in sketches can be recovered by natural language descriptions.

After exploring the modelling of deictic and iconic elements separately, I address the task of interpreting co-verbal gestures, which contain both deictic and iconic elements. As aforementioned, representational gestures bear close temporal and semantic relations to accompanied verbal content. Thus, they do not receive coherent interpretations on their own. Their interpretation must be resolved by reasoning about how they are related to its accompanied verbal content. In this dissertation, I represent the interpretation of verbal content and gestures with a set semantic concepts such as size and shape. Based on an existing corpus of route giving descriptions (i.e., the *SAGA* corpus), I frame the task of representing multimodal descriptions with semantic concepts as a multi-label classification problem. Verbal utterances and hand gestures features are used to learn to predict the semantic categories of co-verbal gestures. I show that natural language is informative for predicting the semantic categories of hand gestures and verbal utterances.

The contributions of this thesis are summarised as follows:

- Two multimodal corpora were collected and publicly available to further research works. The corpora go beyond previous works which either only contain uni-modal data or only include gesture commands rather than multimodal descriptions.

¹<https://www.crowdfower.com/>

- With empirical experiments, this dissertation shows that natural language is informative for interpreting the semantics of accompanied iconic gestures.
- Three variants of representing multimodal descriptions in real-time systems are explored in this dissertation. The results show that automatically learned symbolic labels outperforms verbatim represents and overcome the limitations of representation method with pre-defined symbolic labels.
- This dissertation describes a real-time system which builds and applies multimodal spatial scene descriptions – fusing abstract deictics with speech. The results demonstrate that deictic gestures not only improve overall performance of the spatial interpretation task, but also result in earlier final correct decision of the system due to its parallel nature to language.
- This dissertation investigates the interplay of semantics between natural language and iconic information in sketches, drawing the conclusion that multimodal object descriptions outperform language-only or sketch-only descriptions in an image retrieving task.

1.2 Thesis outline

This thesis is structured as follows:

- Chapter 2 gives an overview of previous work related to the dissertation. Firstly, I introduce previous work on gesture studies which inspect the relation between speech and gestures in natural communications. Secondly, I provide an overview of existing theories on multimodal semantic models and discuss formal semantic representations of speech and co-verbal gestures. Thirdly, I summarise works on multimodal human-computer interfaces, which mainly focus on frameworks and methods of interpreting speech and gesture inputs from humans. I finish this chapter with a discussion of how HCIs can be improved by jointly interpreting natural language and co-verbal gestures/sketches.
- Chapter 3 introduces following multimodal corpora: a) spontaneous spatial scene description corpus. This corpus is composed of intuitive natural language and deictic/iconic hand gestures of a scene description task. With this corpus, I investigate the natural behaviour of spatial scene descriptions and how well natural deictic gestures can represent the spatial configurations in human mind; b) spatial scene descriptions with explicit instructions. This corpus is elicited with a spatial description task similar to previous corpus, however, to collect data amenable for modelling the interpretation of such descriptions with computational methods (Chapter 5), I constrained the setup by making

task-oriented instructions. It results in a corpus with larger amount of multimodal descriptions; c) multimodal object description corpus. In this corpus, real-life photographs are paired with hand-drawn sketches from an existing corpus (Sangkloy et al., 2016) and natural language descriptions collected using Crowdfunder, a crowd-sourcing platform.² This corpus provided materials of investigating symbolic and iconic semantics of objects descriptions in Chapter 6. For each of the corpus, I describe the data collection procedures as well as data statistics.

- Chapter 4 presents three methods of representing multimodal scene descriptions in a computer system. Namely, verbatim representation, representation with pre-defined concepts, and representation with a set of concepts learned from the data. After introducing each method, I describe the evaluation setup and corresponding results, then discuss the pros and cons of each method.
- Chapter 5 presents a real-time system that models the building and application of spatial scene descriptions. The system is supposed to take speech and abstract deictic gestures as input, build representations of the multimodal descriptions and apply the representations to retrieve the target scenes from a set of distractor scenes. First of all, I describe the system framework which is composed of following components: automatic speech recognition (ASR), natural language processing (NLU) module, gesture detection module, gesture interpretation module, multimodal fusion and application module. Then I introduce individual system components and discuss evaluation results of the system performance which demonstrate that deictic gestures not only benefit the overall performance of the system, but also result in earlier final correct decisions .
- Chapter 6 presents a study of investigating the contributions of symbolic and iconic semantic modes in object descriptions. I conduct the investigation with an image retrieving task that takes joint words and hand-drawn sketches as input. After briefly introducing the image retrieving task, I describe the models of grounding words and sketches to images, which judge the fitness between an image and giving words/sketches. Then, I describe how we evaluate the contributions of words and sketches with controlled input from words and sketches, namely, mono-modal and multi-modal experiments. Finally, I discuss the evaluation results and conclude that the iconic information in sketches complement natural language descriptions. This chapter draws a conclusion that even incorporating iconic information from reduced sketches leads to better performance of an image retrieving task.

²www.crowdfunder.com

- Chapter 7 intends to address the task of interpreting co-verbal iconic gestures and construct multimodal representations with a set of semantic concepts. I frame the task of learning multimodal semantic concepts as a multi-label classification task using words and annotations of hand gestures as features. The evaluation results show that natural language is informative for learning the categories of semantic concepts of hand gestures in route giving descriptions.
- Chapter 8 finishes this dissertation with a summarisation of the presented work. This is followed by a discussion on future work of interpreting multimodal communication and building multimodal human-computer interfaces.

In the rest of the dissertation, when referring to work that is my own, I will use *I*, while mentioning work that has been done in collaboration with my co-authors such as experiment design in the studies, I will use *we*.

2

Related work

In this chapter, I introduce background knowledge on human multimodal communication and multimodal human-computer interfaces. This includes an overview of language-related multimodal communications, previous works on hand gestures, temporal and semantic relations between co-verbal gestures and accompanied verbal content. The works in this dissertation benefited from studies in hand gestures. The knowledge in gesture studies forms theoretical hypothesis for building multimodal systems that can understand human multimodal communication. In addition to co-verbal hand gestures, I also mentioned hand-drawn sketches in multimodal communication, as pen inputs are also one of the important input modality in human-computer systems and share some similarities with gestures in terms of conveying iconic information. After discussing previous work in gesture studies, I give an overview of state-of-the-art of multimodal human-computer systems and components of these systems.

2.1 Speech and gestures in natural communications

“We think, therefore, we gesture” (Alibali et al., 2000). When talking, humans often accompany their speech with hand or arm movements. These movements, though different from speech, are part of our communication system that convey meanings together with speech (Quek et al., 2002).

Stated of the art gesture studies show that co-verbal gestures are part of our thinking procedure (Alibali et al., 2000). Co-verbal gestures not only enrich verbal content with useful information, they also help humans to speak (Goldin-Meadow, 2005; Kita, 2000). Kita (2000) proposed the Information Packing Hypothesis. It suggests that gestures conceptualising information for speaking. Hostetter et al. (2007) supports the hypothesis with a study of an ambiguous dot-patter description task. The study shows that participants gesture more frequently when dots are not connected by geometric shapes, suggesting that gestures occur when information is difficult to conceptualise. Moreover, studies also have shown that, even when the verbal content does not match the same spatial ideas in the accompanying speech, representational gestures resemble underlying mental representations Church and Goldin-Meadow (1986); Roth (2002).

In spatial description tasks, humans often produce representational gestures to depict the image they are describing (McNeill, 1992), which also provides a good test case for multi-modal systems (Cassell et al., 2007; Striegnitz et al., 2005; Kopp et al., 2004). Route giving description is such a typical scenario of spatial descriptions, which typically involve verbal descriptions and hand gestures. The most common gesture in route descriptions is pointing gesture, which indicate a direction to follow or direct a listener's attention to a visible landmark from the situated environment. Moreover, in route giving descriptions, people often talk about landmarks and routes that are not in the shared environment. To demonstrate the spatial relation between several landmarks that are not visible, humans often place their hands in the space to represent these landmarks. These "placing gestures", conventionally referred as *abstract deictics* create abstract concepts of the landmarks in the shared space. They map the spatial layout from a speakers' mind to the gesture space, so that a listener can try to imagine the layout in his mind and understand the spatial relations even when the landmarks are not actually visible.

Iconic gestures also often appear in route descriptions, especially when describing shapes of a complex route or landmarks (Cassell et al., 2007; Beattie and Shovelton, 1999; Emmorey et al., 2000b). For example, to clearly describe the route with several turns, a route giver may draw in gesture space to visualise the directions; to specify the contour of a building while describing its colour, name and other attributes with verbal utterances, one might draw the most salient part of the contour e.g., "a dark church with a round window like this [drawing the shape of the window]" or "an elephant with trunk coiled like this [drawing the shape of the trunk]". The descriptions which intends to refer to landmarks are conventionally referred as *Referring expressions*.

Although it's widely accepted that hand gestures do convey meanings, human conversations are rarely composed of pure gestures. This lies in the fact that, in natural conversations, speech

and co-verbal gestures often closely related to each other both on the semantic and temporal level. Hence, the interpretation of co-verbal gestures not only depends its own, but also depends on the coordinated verbal content.

The meaning of hand gestures is also multi-dimensional. For example, an iconic gesture can indicate the size of a window may indicate its shape at the same time. While semantics of gestures concern the meanings, gestures can also function pragmatically such as indicating of emotions (Freigang and Kopp, 2016; Freigang et al., 2017). In this dissertation, I focus on the semantics of iconic gestures without considering the pragmatics.

Pen-input such as *lines* and *circles* are also commonly used in route giving descriptions when a sketch board or a piece of paper is available. For example, a route giver can circle a landmark on a map to indicate the selection of that location; he/she can also draw a short line to indicate the *direction*, or even several connected lines to signify a route a listener should follow Bolt (1998); Hui and Meng (2014). Thus, these pen inputs' functions are similar to pointing gestures in the sense of intending to locating landmarks or giving directions.

Similar to iconic gestures, pen inputs can also enrich verbal utterances with iconic information when giving descriptions such as drawing the contour of a building or an object, e.g., drawing a landmark to visually signify its shape. However, as aforementioned, pen-inputs are also different from iconic gesture as sketches can encode much more details than iconic gestures. Moreover, sketches with full details can convey information on their own, hence, in these cases, sketches are only loosely related to the accompanying speech both on the semantic and temporal level, e.g., a full sketch of a cat is informative as a depiction of the cat, verbal descriptions doesn't have to co-occur with the sketch to make it informative such as in sketch-based image retrieval tasks (Eitz et al., 2011; Li et al., 2012; Sangkloy et al., 2016).

Despite the fact that pen-inputs are able to enrich verbal content, it requires support of devices such as pens, papers or sketch board to perform the functions. In comparison, hand gestures, without requiring support from other devices, appear more often and more natural in situated conversations. In what follows, I will first have a look at the typologies of hand gestures in natural communications, then discuss representational gestures: abstract deictic gestures and iconic gestures, as well as the semantic and temporal relations between co-verbal gestures and the accompanying speech.

2.2 Typologies of hand gestures

In this section, I give an overview of gesture categories based on gesture movements and the relation between gestures and the accompanied speech. Although this dissertation only concerns representational gestures, to given a complete view of hand gestures in natural communica-

tions, I review all categories of gestures while focusing the discussions on representational gestures.

According to the characteristics of gesture movements, hand gestures are usually categorised as: *iconic*, *deictic*, *metaphoric*, and *beats* (McNeill, 1992).

- **Iconic gestures** represent concrete objects by resembling the visual similarities between gestures and the referred objects. For example, drawing in the space to indicate the shape of a window. Hence, they bear close formal relationship to the semantic content of the verbal utterances.
- **Deictics** are also referred as *pointing gestures* which often communicate by directing a listener's attention to the spot it points to. Such deictic gestures are featured with the index finger extended, other fingers closing. However, much of the deictics we see in daily conversations are actually abstract deictics which do not point to visible objects in the situated environment and are not with extended index finger, but point to the space to create an imagined (abstract) object in the shared environment (McNeill, 2005).
- **metaphoric gestures** present an image of an abstract concept such as knowledge, thus, metaphoric gestures often indicate that the accompanying speech is meta, rather than concrete objects.
- **beats** are movements which do not present discernible meanings, but can be recognised by the pattern of their movements. Beats can function to signal the temporal locus of something a speaker thinks important. That is, to stress the important of something.

In this dissertation, I focus on deictics in spatial descriptions and iconic gestures in route descriptions.

Kendon's continuum According to the relation between gestures and the accompanied speech, **Kendon's continuum** (McNeill, 1992) distinguishes gestures of different kinds along as continuum as shown in Figure 2.1. Along the continuum from left to right, two kinds of reciprocal changes occur: the degree of semantic relations between speech and gestures decreases, while the degree of a gesture shows the properties of a language increases.

- **Gesticulation** is the most frequent type of gestures in our daily communications. It refers to gestures that embody meanings relatable to the accompanied speech, e.g., iconic gestures and abstract deictic gestures. Therefore, gesticulation bears close semantic and temporal relations with the accompanied speech. The stroke phase of gesticulations often precede or synchronise the accompanying speech (Kendon, 1980a).


Gesticulation Speech-framed-gestures Emblems Pantomime Signs 

Figure 2.1: Kendon’s Continuum. As one moves from left to right, the degree of semantic relations between speech and gestures decreases, while the degree of a gesture shows the properties of a language increases.

- **Speech-framed-gestures** can be considered as part of the accompanied speech. As a result, speech-framed-gestures do not synchronise with the accompanied speech, but fill grammar slots. McNeill (2006) gives an example of speech-framed-gestures as follows: “Sylvester went [gesture of an object flying out laterally]”.
- **Emblems** are also referred as symbolic gestures. Emblems are gestures with conventional meanings, e.g., thumbs-up for “great”. The meaning of emblems may vary across different cultures.
- **Pantomime** can be one or a sequence of gestures that tell a story, produced without speech. Pantomime is also referred as dumbshow.
- **Signs** are a different language such as ASL. Each sign functions as a lexical in sign language, thus, the least relevant to the accompanying speech.

The gestures this dissertation concerns fall into the gesticulation category and closely related to the accompanying speech.

2.2.1 Abstract deictics in spatial descriptions

Gestures are not limited to describe concrete world, they can also describe objects that are not in the situated environment. e.g., objects out of current view and the relations between them. In route giving descriptions, abstract deictics are often deployed to exhibit spatial layout of landmarks (Cassell et al., 2007). In such cases, anchoring the destination in configurations of landmarks and indicating their relative spatial layout with deictic gestures pointing into the empty gestural space is a common practice (Emmorey et al., 2000a; Alibali, 2005; Cassell et al., 2007).

In multimodal route descriptions, deictic gestures map the spatial layout of the landmarks from the speaker’s mental image to the shared gesture space (McNeill, 1992). Together with the verbal descriptions, a listener can build a mental representation of the landmarks, later navigating itself by comparing the mental representation with real-world landmarks.

While the verbal utterances describe some important attributes of the referential objects (e.g., entity name: *the bus stop*, relative position: *a bit left of*), the deictic gestures complement the verbal content with spatial information (i.e., points with coordinates in the gesture space, standing in for the real locations of the referents, and indicating their spatial relation). When combining the verbal content with gestures, a listener may form a complete and more accurate understanding of the description (e.g., how much *left* is *a bit left*, relative to *below*). Importantly, such deictic gestures only encode position information, thus their meanings rely on the temporally aligned verbal content. That is, an abstract deictic is meaningful when the accompanied verbal content describes other attributes of an object, otherwise, the deictic would not receive a defined meaning. Hence, the task of interpreting such descriptions goes beyond previous works on pointing gestures, in which gestures can be grounded to objects present in the environment (Stiefelhagen et al., 2004).

Psycholinguistic studies show that humans process gestures and speech jointly and incrementally (Campana et al., 2005). While descriptions unfold, listeners immediately integrate information from co-occurring speech and gestures. Moreover, to apply the interpretation later, it's essential to form a hypothesis in mind, making it a very demanding cognitive, language-related task (Schneider and Taylor, 1999). Hence, incremental processing is essential to build a real-time system that can understand the descriptions in the way humans do (Schlangen and Skantze, 2009).

2.2.2 Describe objects with iconic gestures/sketches

Humans often use iconic gestures to describe object (i.e., referential expression), and iconic gestures are convenient to convey visual information which might be difficult to encode in language (McNeill, 1992).

For example, one can describe an elephant with the utterance “*an elephant facing right trunk coiled towards mouth*”. While the utterance gives accurate information the category of the entity (i.e., *elephant*), it does not specify the exact shape of the trunk. Consequently, a listener's mental representation of the “trunk” is ambiguous. Accompanying the utterance with an iconic gesture that draws the shape of the trunk may help a listener to understand the description with more accurate details.

Due to the nature that iconic gestures convey meanings by resembling visual similarities, the same iconic gestures when accompanied with different verbal content, can convey different meanings. For example, an iconic gesture with a coiled trajectory may indicate the shape of an elephant in our previous example, it can also indicate locations when accompanied by the utterance: “from A to B” (McNeill, 1992; Sowa and Wachsmuth, 2003).

Especially descriptions of visual objects or situations can be supported by the iconic mode

of reference provided by gestures or sketches, that is, reference via similarity rather than via symbolic convention (Pierce, 1867; Kendon, 1980b; McNeill, 1992; Beattie and Shovelton, 1999).

2.3 Relations between speech and co-verbal hand gestures

In this section, I discuss previous studies on the relations between co-verbal gestures and accompanied verbal content both on semantic and temporal level.

2.3.1 Semantic coordinations between co-verbal gestures and verbal content

In the Growth Points in thinking for speaking model, McNeill and Duncan (1998) claims that speech and gestures are systematically organised in relation to one another, although they express the same underlying ideas, but in different modalities and not necessarily express identical aspects of the ideas. In many cases, the two modalities serve to reinforce one another, e.g., the drawing gesture of an elephant's trunk enriches the verbal description with shape information that are not exactly covered in language. In such cases, the information to be expressed is distributed across both modalities such that the full communicative intentions of the speaker are interpreted by combining verbal and gestural information. The semantic synchrony of both modalities can be thought of as a continuum of co-expressivity, with gestures encoding completely the same aspects of meaning as speech on one extreme (Bergmann et al., 2011; Kita and Özyürek, 2003).

When speech and gestures express the same meanings, gestures may seem to be redundant in the descriptions. For example, humans may describe a fountain as “round” while drawing a circle to indicate the shape of the fountain. In this case, the drawing gesture does not add extra information to enrich the verbal content, but visualises the same information so that a listener can “see” the shape of the fountain.

In the two above example, the information in iconic gestures were also expressed by the verbal content, either partially or completely. Iconic gestures can also encode information that are not uttered verbally. That is, these gestures complement speech. For example, one can describe a fountain with utterance “a fountain” while drawing a circle to indicate its shape. Without the accompanied gesture, a listener's mental representation of the fountain would miss the shape information. Only when combining both modalities, a listener can form a more complete representation (Pine et al., 2007; McNeill, 1992).

Although the semantic coordination between speech and co-verbal gestures have been used to generate speech and co-verbal gestures (Kopp and Bergmann, 2017b; Bergmann et al.,

2013a), human-computer interfaces rarely deploy this knowledge to interpret multimodal communications, but focus on the semantics of human gestures with conventions or pre-defined rules.

2.3.2 Temporal alignment between gestures and speech

Besides close semantic coordinations, speech and co-verbal gestures also bear close temporal relations. As verbal utterances unfold word-by-word in situated conversations, co-verbal gestures often co-occur with the words they shared the same semantic meanings (Nobe, 2000; Schegloff, 1984; McNeill, 1992, 2005).

Studies have shown that when talking, speakers produce a perceptible link between the motion they impose upon a referent and the prosodic structure of their speech. Listeners readily use this prosodic cross-modal relationship to resolve referential ambiguity in word-learning situations (Jesse and Johnson, 2012; Özyürek et al., 2007). Temporally unaligned gestures and speech often result in mis-understandings of the content.

Chui (2005) found, in Chinese, there is a higher proportion of gestures synchronised with speech than gestures anticipating speech. In English, on the contrary, Schegloff (1984) observed that gesture strokes are generally produced in anticipation to lexical affiliates. Similarly, Leonard and Cummins (2009) also found an anticipation of gestures in English.

Although temporal and semantic relations between speech and gestures are not indecent, they affect each other. For example, Bergmann et al. (2011) investigated how far temporal synchrony is affected by the semantic relationship of gestures and their lexical affiliates in the SAGA corpus (Lücking et al., 2010). The results showed that when gestures encode redundant information, gestures' onsets are closer to that of co-occurred lexical affiliates than when gestures convey complementary information. That is, the closer speech and gestures are related semantically, the closer is their temporal relation.

By far, I have had an overview of gestures in natural communications, categories of hand gestures, as well as the relation between representational gestures and accompanied speech. Next, I provide an overview of state-of-the-art multimodal interfaces which are designed to interpret multimodal communication and respond to multimodal input from humans.

2.4 Multimodal human-computer interfaces

In this section, I will first give an overview of previous works on general frameworks of multimodal interfaces, then discuss works on individual components of M-HCIs such as natural language processing, gesture recognition and multimodal fusion.

Multimodal human-computer interfaces aim to enable computers/robots/virtual agents to understand multimodal human communications in the way humans do. Therefore, a M-HCI must be able to understand natural language (NLU), recognise and interpret hand gestures, combine information from both modalities (multimodal fusion) and represent multimodal content in a way that later can be applied to real-life tasks with humans, e.g., after hearing a route giving description, a robot should be able to navigate itself. In other words, a multimodal system is usually composed of two pipelines: a natural language processing pipeline and a gesture processing pipeline. A fusion engine takes the outputs from the two pipelines and form a joint interpretation of multimodal input (Oviatt and Cohen, 2000; Oviatt, 2003; Dumas et al., 2009; Turk, 2014).

Since the seminal work of Bolt (1998), prominent progresses have been made on advancing machines' ability to understand multimodal communication from humans. Most of the early works are on concept demonstration level without building computational models. For example, Koons et al. (1993b) describes two prototype systems that accept simultaneous speech, gestural and eye movement input. The task of the systems was to resolve objects in a map by processing the three modes to a common frame-based encoding (representation) and interpreting the encoding. Similar to Bolt (1998), as the two systems are only prototypes, natural language processing and gesture processing methods were not described. Koons et al. (1993a) discussed the integration of information from speech, gestures, and gaze at computer interfaces. Two prototype systems were proposed, where speech, gestures and eye gaze are processed to a common frame-based encoding and interpreted together to resolve references to objects in a map. Cohen et al. (1997) describes an agent-based, collaborative multimodal system - the Quick system. Quick enables a user to create and position entities on a map or virtual terrain with speech, pen-based gestures, and/or direction manipulation. Cassell et al. (1999) introduced an embodied conversational agent that is able to interpret multimodal input and generate multimodal output. Although the input gestures are limited to "giving turn". Chai et al. (2002) presents a semantics-based multimodal interpretation framework - Multimodal Interpretation for Natural Dialog (MIND). The system can take graphics, speech and video inputs for simple conversations with humans.

Recent years have seen fast development of high resolution cameras that are widely deployed to record audio and video data and infrared devices used for tracking body movements such as Kinect and Leap sensor. These advancements enable research on human-computer interfaces that can understand multimodal human communication. As a result, multimodal systems started to go beyond prototypes and concept demonstrations. For example, Zhu et al. (2002) proposed a real-time multimodal system to spot, represent and recognise hand gestures from a video stream. Johnston et al. (2002) describes MATCH, a multimodal applica-

tion architecture that combines finite-state multimodal language processing, a speech-act based multimodal dialogue manager, dynamic multimodal output generation, and user-tailored text planning to enable rapid prototyping of multimodal interfaces with flexible input and adaptive output. A gesture and handwriting recogniser provides possible classifications of 285 words and a set of 10 basic gestures such as *lines*, *arrows* and *areas*. Nickel and Stiefelhagen (2003) presented a system capable of visually detecting pointing gestures and estimating the 3D pointing direction in real-time.

Hoste et al. (2011) introduced Mura, an integrated multimodal interaction framework. The framework supports the integrated processing of low-level data streams as well as high-level semantic inferences to fully exploit the power of multimodal interactions. However, it didn't address the interpretation and semantic representation of iconic gestures, but merely a concept demonstration. Lucignano et al. (2013) presented a POMDP-based dialogue system for multimodal human-robot interaction. The system is able to recognise 9 gestures with possible meanings, each of which is with a specified interpretation, e.g., a *hand's palm stop* gesture for "stop, stop down". Matuszek et al. (2014) demonstrate that combining unscripted deictic gestures and verbal utterances more effectively captures user intent of referring to objects in human-robot interactions. Whitney et al. (2016) defined a multimodal Bayes filter to interpret a person's referential expressions to objects. The approach incorporated learned contextual dependencies composed of words and pointing gestures. Hui and Meng (2014) describes an approach in semantic interpretation of speech and pen input using latent semantic analysis (LSA) in the navigation domain. The pen inputs can be categorised as point (indicate a single location), circle (small one indicate a single location; larger one indicate multiple locations) and stroke (indicate either a single location or start and end points of a route).

McGuire et al. (2002) reported progress in building a hybrid architecture that combines statistical methods, neural networks, and finite state machines into an integrated system for instructing grasping tasks by man-machine interaction. The system combines the GRAVIS-robot for visual attention and gestural instruction with an intelligent interface for speech recognition and linguistic interpretation, and a modality fusion module to allow multi-modal task-oriented man-machine communication with respect to dextrous robot manipulation of objects with 3-D pointing projection.

To summarise, existing multimodal systems are designed to take pointing gestures, symbolic gestures and a set of gestures/pen input commands as input. Although these systems have various system architecture, some of them designed to take various types of input modality, unfortunately these systems are only able to interpret pre-specified gesture inputs, a subset of gestures/pen input of natural communication. To interact with such systems, a user have to remember the patterns and meanings of the gesture commands (e.g., a circle for "selecting").

Thus the interaction between systems and users is far from natural communication.

Moreover, previous works on multimodal systems rarely considers speech and gestures communication as time sequence inputs. The processing of multimodal inputs were on gesture-speech unit level. In other words, only at the end of a gesture-speech input, the processing of the input starts. An important issue with this method is that the temporal relations between speech and gestures are ignored.

There is some disagreement among researchers about the role of gesture in comprehension; whether it is ignored, processed separately from speech, used only when speakers are having difficulty, or immediately integrated with the content of the cooccurring speech. Campana et al. (2005) presented an experiment that provides evidence in support of immediate integration. In the experiment, participants watched videos of a woman describing simple shapes on a display in which the video was surrounded by four potential referents: the target, a speech competitor, a gesture competitor, and an unrelated foil. The task was to “click on the shape that the speaker was describing”. In half of the videos the speaker used a natural combination of speech and gesture. In the other half, the speaker’s hands remained in her lap. Reaction time and eye-movement data from this experiment provide a strong demonstration that as an utterance unfolds, listeners immediately integrate information from naturally cooccurring speech and gesture.

In this dissertation, I consider incremental processing for speech and gesture inputs and deploy the temporal relations between speech and gestures to enable a multimodal system achieve earlier correct decisions (Han et al., 2018) (see Chapter 5 for details).

After having an overview of previous work on multimodal systems, in the rest of this section, I will discuss individual modules that compose multimodal systems, namely, natural language processing module, gesture recognition and interpretation module, multimodal fusion module and the multimodal representation module.

2.4.1 Natural language processing

The natural language processing pipeline in a multimodal system takes verbal utterances as input, and provides the fusion module with certain representations of the verbal content to be fused with other modalities.

As noted by Roy and Reiter (2005), language is never used in isolation; the meanings of words are learned based on how they are used in contexts—for the spatial description task of this dissertation, *visual* contexts—where visually-perceivable scenes are described (albeit scenes that are later visually perceived). This approach to semantics is known as *grounding*; previous works such as (Gorniak and Roy, 2004, 2005; Reckman et al., 2010) discussed how word meanings such as colour, shape, and spatial terms were learned by resolving referring

expressions. (Harnad, 1990; Steels and Kaplan, 1999) observed that symbolic approaches to semantic meaning (e.g., first-order logic) do not model such perceptual word meanings well; Harnad (1990) and Larsson (2013) reconcile grounded semantics and symbolic approaches. In this dissertation, I extend earlier work in this area (Larsson, 2013; Kennington et al., 2015) by learning and applying these mappings in a spatial description and scene retrieval task.

Navigation tasks provide a natural environment for the development and application of such a model of grounded semantics, which have been the subject of a fair amount of recent research: In (Levit and Roy, 2007), later extended in (Kollar et al., 2010), the meaning of words related to map-navigation such as “toward” and “between” were learned from interaction data. Vogel and Jurafsky (2010) applied reinforcement learning to the task of learning the mapping between words in direction descriptions and routes. Also, Artzi and Zettlemoyer (2013) learned a semantic abstraction from the interaction map-task data in the form of a combinatory categorical grammar. Though interesting in their own right, these tasks made some important simplifying assumptions that we go beyond in this paper: first, gestural information was never used to convey scene descriptions; second, the scene that is being described (from a bird’s-eye view; here, scenes are perceived from a first-person perspective) was visually-present at the time the descriptions are being made; third, only the grounded semantics of a selected subset of words were being learned. In this dissertation, gestures are considered, a description is heard and *later* applied to scene retrieval tasks, and all the word groundings are learned from data.

Kintsch and van Dijk (1978) suggested that listeners first represent exact words of a description (i.e., surface form), then interpret information (i.e., a *gist* of the description) and integrate that with their world knowledge (e.g., the knowledge about what red things look like, if the word “red” was used in the description). Moreover, Brunyé and Taylor (2008) (as well as some work cited there) note that readers construct cohesive mental models of what a text describes, integrating time, space, causality, intention, and person- and object-related information. That is, readers progress beyond the text itself to represent the described situation; detailed information from an instruction or description is distorted in memory (Moar and Bower, 1983), thus, incremental processing is essential in processing such descriptions.

2.4.2 Gesture recognition and interpretation

As aforementioned, current multimodal systems mainly concern gestures/pen input of certain patterns, each of which with specified meanings. In these systems, a gesture classifier recognises input gestures as one of the gestures in the pre-defined gesture set; a gesture interpretation module maps input gestures to actions/decisions the system should make according to the pre-defined gesture-action/decision map. For example, when a “thumb up” gesture is detected, an interactive system may consider it as a confirmation signal from a user, and consequently

end the current interaction session. Next, we first given an overview on gesture recognition methods, then discuss state-of-the-art gesture interpretation modules in multimodal systems.

Gesture recognition Depending on the data types a multimodal system receives from its sensors, gesture recognition methods can be categorised as video-based gesture recognition (Wu and Huang, 1999; Murthy and Jadon, 2009; Rautaray and Agrawal, 2015) or motion-based gesture recognition (Bayazit et al., 2009; Wu et al., 2013; Marin et al., 2014; Pitsikalis et al., 2017). Approaches from the two categories differ in the way they extract features from the raw data. Motion data often provides rich skeleton details of hands or limbs such as joint positions in each data frame of a gesture, making it easier to extract high level features to represent gestures. In comparison, video-based gesture recognition methods often have to re-construct the motion information from videos (a difficult task on its own), making the recognition task more challenging than motion-based recognitions.

Depending on the characteristics of the gestures, gesture recognition methods can be categorised into static gesture recognition (Hasan and Abdul-Kareem, 2014) and dynamic gesture recognition (Joslin et al., 2005). While the former one is for static gestures such as “OK” (Freeman and Roth, 1995; Hasan and Abdul-Kareem, 2014), the latter approach is for gestures composed of a series of movements such as circling and iconic gestures (e.g., drawing the contour of a vase), where temporal scale of gestures should be considered. For example, Sadeghipour and Kopp (2014) proposed the framework of FSCFG, which combines feature-based representation with syntactical rule-based organisation to learn a grammar of natural iconic gestures.

Recurrent neural networks (RNN) (Gers et al., 1999) have also become popular for gesture recognition tasks as it considers temporal relations among series data via a memory mechanism (Molchanov et al., 2015; Wu et al., 2016). Although deep neural networks have achieved impressive performances on gesture recognition tasks, training such models require large amount of data which might be unavailable. In this dissertation, I introduce a gesture recogniser to detect abstract deictics that adopts a long-short-memory network (LSTM) Han et al. (2018) (See Chapter 5 for more details).

Gesture interpretation Gesture interpretation modules inform the system what decision should be made when a gesture is detected. For gestures with pre-defined meanings, the interpretation module simply maps the gesture to corresponding decisions defined in the system or produces probabilities over all decisions which indicate how likely a gesture is meant for each decision. As this strategy relies on a pre-defined map between gestures and system decisions, the systems are only able to handle a limited set of gesture input.

As the meaning of representational gestures/sketches relate to the accompanying speech, the interpretation of representational gestures is more complex. For example, the interpretation of iconic gestures should represent the encoded iconic information. Deep neural networks have been effective at encoding visual features such as real-life photos (Simonyan and Zisserman, 2015). Recently, DNNs have been applied to image classification and sketch-based image retrieval tasks by encoding images as feature vectors. The feature vectors encoded by the neural networks have shown good performance on other tasks for representing images (Koch et al., 2015; Collell Talleda and Moens, 2016).

In this dissertation, due to the lack of data, I only consider interpreting and representing sketches as vectors with the GoogLe network, which was originally introduced in (Sangkloy et al., 2016), leaving the interpretation of iconic gestures as future work (see Chapter 6 for details).

2.4.3 Multimodal fusion

A key step of modelling the interpretation of multimodal communication is to combine information from individual modalities to form a complete understanding of the content. This step is often referred as multimodal fusion or multimodal integration (in the rest of the dissertation, I will use the term multimodal fusion).

Existing multimodal fusion approaches can be categorised as non-temporal models such as early fusion, late fusion, and hybrid fusion that combines early fusion and later fusion approaches, or spatial-temporal neural network models that consider temporal relations between individual modalities (Atrey et al., 2010). Table 2.1 shows a summarisation of multimodal fusion approaches and respective characteristics. Next, I give an overview of these fusion approaches and discuss the features and application scenarios of each approach.

- **Early fusion:** The early fusion approach fuses individual modalities at the feature level, thus it is also referred as feature fusion. It assumes all modalities are tightly synchronised on the temporal level. Therefore, the features of different modalities at the same time point are concatenated for the classification task. Early fusion suits for classification tasks that involve temporally synchronised modalities, such as speech and lip movements for speech recognition tasks and acoustic features and linguistic features for emotion recognition tasks (Schuller et al., 2005).
- **Late fusion:** Late fusion approach combines different modalities on the semantic level or decision level, requiring a recogniser or processing module for each modality. Therefore, it is most suitable for modalities that are only loosely synchronised on the temporal level such as speech and hand gestures.

| Approach | fusion |
|------------------|----------------------------------|
| Early fusion | feature level / data level |
| Late fusion | decision level |
| Hybrid fusion | feature level + decision level |
| Attention models | feature level+temporal relations |

Table 2.1: Overview of multimodal fusion approaches and respective characteristics.

For example, Johnston et al. (1997) proposed a unification-based approach of multimodal fusion. Integration of spoken and gestural input is driven by unification of typed feature structures representing the semantic contributions of the different modes. Wu et al. (1999) presented a statistical approach to integrate information in modalities. The approach fuses the posterior probabilities of parallel input signals involved in the multimodal system, in which the posterior probabilities were determined by the recognisers of each modality. Lucignano et al. (2013) adopted a late fusion approach to combine verbal commands with gesture actions in a multimodal human-robot interaction dialogue system. Each modality recogniser (i.e., speech and gesture recognisers) provides the fusion engine with a N-best list of possible interpretations, so that the engine can form a joint representation and consequently provides the dialogue manager with a N-best list possible interpretations.

- **Hybrid fusion:** The hybrid fusion approach aims to take the advantage of early fusion and late fusion approaches. Therefore, it is often adopted in cases where multiple modalities are involved for a classification task (Bendjebbour et al., 2001; Xu and Chua, 2006). When multiple modalities are involved while only some of them are with close temporal relations, using either early fusion or late fusion might result in suboptimal classification results for some of the modalities. For example, Snoek et al. (2005) considered both early fusion and late fusion for the task of semantic analysis of multimodal video for 20 semantic concepts. The results showed that the late fusion approach tends to give slightly better performance for most concepts. But, for those concepts where early fusion performs better, the difference is more significant. Thus, a hybrid fusion approach might lead to better classification results.
- **Spatial and attention based neural network models for multimodal fusion:** Recently, deep neural network with spatial and temporal attention models haven also been de-

ployed to perform the fusion task, mainly for sentiment analysis tasks (Chen et al., 2017; Neverova et al., 2016). For example, Zadeh et al. (2018) presents a neural architecture for understanding human communication called the Multi-attention Recurrent Network (MARN). The main strength of the model comes from discovering interactions between modalities through time using a neural component called the Multi-attention Block (MAB) and storing them in the hybrid memory of a recurrent component called the Long-short Term Hybrid Memory (LSTHM). Baltrušaitis et al. (2018) provides an overview of multimodal machine learning which aims to build models that can process and relate information from multiple modalities using neural networks, aiming to go beyond the typical early and late fusion categorisation and identify broader challenges faced by multimodal machine learning, and enable researchers to better understand the state of the field and identify directions for future research.

Fusion engines in multimodal systems should also consider user-adaptive mechanisms to take into account of users' preferences to certain modalities, e.g., gestures/pen input might be less reliable than speech. Epps et al. (2004) analysed data collected from a speech and manual gesture-based digital photo management application scenario, and found for that application, about 37% of tasks were completed using unimodal rather than multimodal input. Hence, the fusion engine should adapt its mechanism according to the input modalities. Chai et al. (2002) used context information to enhance the fusion module to handle inaccurate human inputs.

In this dissertation, I aim to model the interpretation of multimodal spatial descriptions, where abstract deictics and natural language temporally correlate to each other but not tightly synchronised, therefore a late fusion approach is adopted to combine verbal descriptions with gestures/sketches (see Chapter 5 for details).

2.5 Representation of multimodal content

The ultimate goal of building multimodal interfaces is to understand multimodal communication from humans, extract useful information on the semantic level: what has been described? What properties the described object has? As information in multimodal communication often spread across all modalities, a knowledge representation model is required to properly represent the integrated knowledge, which can later be applied to generate system replies or perform tasks (Niekrasz and Purver, 2006).

One of the representation approach of multimodal utterances is the Imagistic Description Theory model (IDT), which was developed based on an empirical study to capture the imagistic content of shape-related gestures in a gesture interpretation system (Sowa and Wachsmuth,

2005; Sowa, 2006). It was designed to capture all meaningful visuo-spatial features in shape-depicting iconic gesture. Each node in an IDT contains an Imagistic Description which holds an object schema representing the shape of an object or object part, respectively (Sowa and Kopp, 2003). Bergmann and Kopp (2008a,b) employ such a hierarchical model of IDT model to represent multimodal chunks for speech and gesture production. Sowa and Wachsmuth (2009) also proposed a unified shape representation for multimodal descriptions involving speech and ionic gestures, which used an *Image Description Tree* (IDT) to conceptualise the concepts in multimodal signals. Each object in the tree system is represented with a set of properties.

Rieser and Poesio (2009) discussed how PTT, a dialogue theory (Poesio and Traum, 1997), can be extended to provide an incremental modelling of speech plus gesture in interactive dialogues where grounding between dialogue participants was obtained through gesture. However, the representation was on the discourse level, but not utterance level.

Lascarides and Stone (2009) provided a formal semantic analysis of co-verbal iconic and deictic gestures. The content of language and gestures was proposed to be represented jointly in the same logical language, where rhetorical relations connect the content of iconic gesture to that of its synchronous speech, and language and gestures are interpreted jointly within an integrated architecture for linking utterance form and meaning. The model exploited discourse structure and dynamic semantics to account for co-reference across speech and gesture and across sequences of gestures.

As for the modelling of spatial descriptions in this dissertation, the spatial descriptions are composed of short multimodal descriptions that include only one to three object descriptions. Such simple descriptions don't include tree structures of object properties which can benefit from the IDT representation method. Due to the short durations of the descriptions, they also don't contain complex discourse phenomena. Hence, instead of using the above mentioned models for discourse representations, we adopted a simple representation structure inspired by the discourse representation theory (DRT) (Kamp and Reyle, 2013; Asher and Lascarides, 2003) is as follows (Han et al., 2015):

$$(1) \quad \begin{array}{l} o_1, g_1, o_2, g_2 \\ o_1: \text{transl}(\text{red circle}) \\ g_1: (x_1, y_1) \\ \quad \text{pos}(o_1, \phi(g_1)) \\ \text{slightly_above}(o_1, o_2) \\ o_2: \text{transl}(\text{blue L}) \\ g_2: (x_2, y_2) \\ \quad \text{pos}(o_2, \phi(g_2)) \end{array}$$

where $\text{transl}(\cdot)$ and $\phi(\cdot)$ indicate functions that translate input utterances and gestures into representations of symbols with values derived from the input signal. Hence, ultimately, we represent each referential object in the descriptions with a set of symbols. These symbols can be words in natural language, visual properties such as colour, shape and size, or even a set of automatically learned symbols without manual specification. In Chapter 4, we explored the three representation variants.

Representing spatial descriptions with symbols also enable us to adopt the “words-as-classifiers” (WAC) model when applying the representations to image retrieval tasks (Kennington and Schlangen, 2015; Schlangen et al., 2016) (See chapter 4 for more details).

2.6 Existing multimodal datasets

Multimodal datasets are essential for building and evaluating computational models of interpreting multimodal communication. With the development of easily available video and audio recording devices such as high resolution cameras, as well as portable motion tracking devices such as Kinect¹ and Leap sensor², the collection of conversational/discourse level datasets has also been facilitated in recent years.

Multimodal corpora of natural multimodal human-human conversations composed of rich natural behaviours are ideal materials for building computational models of HCIs. For example, the *Bielefeld Speech and Gesture Alignment Corpus* (SAGA) introduced a multimodal corpus of 25 route giving and following conversations (Lücking et al., 2010). Besides audio and video recordings of the dialogues, the corpus also provides detailed annotations of hand

¹<https://developer.microsoft.com/en-us/windows/kinect>

²<http://www.leapmotion.com>

gesture features (e.g., palm direction, hand movement), gesture semantics as well as the coordinations between hand gestures and accompanied verbal utterances. The SAGA corpus provides good materials for building and evaluating computational models of construction multimodal semantics with verbal utterances and hand gesture features, which will be described in Chapter 7.

Blache et al. (2009) presented an annotated multimodal corpus of the ToMA project, providing a general framework for building and annotating multimodal corpora that considers phonetics, morphology, syntax, discourse and gestures. While these datasets provide good materials for conversational/discourse level analysis of multimodal communication, it is not publicly available for further research work.

To date, publicly available multimodal corpora are mainly for modelling multimodal human-machine interactions. Hence, these corpora often include multimodal commands, rather than multimodal human-human conversations. For example, Schiel et al. (2002) presented a corpus of human-machine communication combining acoustic, visual and tactile input and output modalities. 90 session recordings of 45 users (100 volumes, DVD-5 format) were distributed with the basic distribution costs of 255 Euro per volume. Fotinea et al. (2016) presented the MOBOT dataset, a dataset of multimodal commands which includes speech and motion data of human limb movements. Kousidis et al. (2013a) presented the TAKE corpus which includes gaze, pointing gestures and verbal utterances. The data was elicited with a Wizard-of-Oz scenario where participants instructed a “system” to choose pentomino pieces from a screen. The data includes 20 gestural commands such as *Stop sign*. Although these corpora contain natural multimodal communication behaviours, the gestures in these corpora are either emblem gestures with conventional meanings or pointing gestures, whose interpretations are not relatable to accompanied speech.

Besides multimodal corpora of video/audio recordings, there are publicly available 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), the gestures in these corpora are often emblems rather than gesticulations. Moreover, most of these existing datasets are collected for gesture classification tasks without accompanied speech/verbal utterances.

To bridge the gap, I collected a corpus of multimodal spatial scene descriptions, including abstract deictic gestures and verbal descriptions of simple scenes. The data was elicited with a simple task in which multimodal descriptions were jointly used to give scene descriptions. Video and audio recordings as well as natural hand motion data were recorded. I will introduce the corpus in Chapter 3.

To model the interpretation of multimodal descriptions involving natural language and

iconic information from pen inputs, I augmented an existing dataset - the Sketchy dataset - with verbal descriptions Sangkloy et al. (2016). The Sketchy dataset provides hand-drawn sketches of objects in real-life photos, which signifies the iconic aspects of real-life objects. For each of the photo-sketch pairs, I collected verbal descriptions that describe visual attributes of objects such as colour, orientation, materials and so on. Therefore, the verbal utterances and the sketches jointly forms multimodal descriptions of real-life objects. Detailed description of the corpus will be provided in Chapter 3.

2.7 Summary

In this chapter, I overviewed and discussed previous work on multimodal communication related to co-verbal hand gestures. The discussion was organised into: hand gestures in natural human communication and the taxonomy of co-verbal gestures; the semantic and temporal relations between verbal and hand gesture content. I also discussed how to compute multimodal semantics, mostly focusing on computational models. Finally, I provide an overview of multimodal interfaces which are designed to understand speech and co-verbal hand gestures in human communication.

3

Multimodal corpora

In this chapter, I introduce three multimodal corpora that are related to studies in subsequent chapters of this dissertation. I start by introducing the *Multimodal Spatial Scene Description Corpus*, in which natural language and abstract deictic gestures are deployed to describe spatial scene images, then describe the *Multimodal Object Description Corpus*, where natural language and hand-drawn sketches jointly describe objects in real-life photographs. Finally, I briefly describe the *Bielefeld Speech and Gesture Alignment Corpus (SAGA)*, which was originally introduced in Lücking et al. (2010).

3.1 Multimodal spatial scene description corpus

Route description, a common task in our daily life, is typically performed with speech and hand gestures (Emmorey et al., 2000a). When describing routes that are not visible in the shared environment, humans often anchor the target place to surrounded landmarks by specifying their relative positions to each other. Visual attributes of these landmarks such as colour, entity name are often described verbally, while the contour of landmarks and trajectories of the routes can be conveniently demonstrated with hand gestures (Alibali, 2005; Cassell et al., 2007). For example, when helping a person to locate a hotel not in current view, the route giver might think of a few landmarks around the hotel as shown in Figure 3.1, and describe the route as



Figure 3.1: Spatial layout of landmarks in Example (1).

follows:

- (1) You take the subway and get off at the Market Street station. Here_[deictic] is the station. Here_[deictic] is a fountain. If you walk _[iconic] around it, you will see the hotel _[deictic] here.

While verbal descriptions specify the names of the landmarks and actions (i.e., *station, a bit left* and *walk around*), the deictic gestures demonstrate the spatial layout of the landmarks with hand positions; the iconic gesture visually signifies the trajectory of the route. Only by combining the verbal descriptions and the hand gestures, it's possible to form a complete and accurate interpretation of the description. To form such a complete interpretation and later navigate itself, the listener must pay attention to verbal descriptions and hand gestures at the same time, interpret the verbal descriptions and hand gestures in parallel while the description unfolds, fuse information from both modalities and represent the the interpretation in his/her mind, making it a challenging task even for humans (Schneider and Taylor, 1999). The route giving and following task forms a good test case of situated dialogue understanding, as the the accuracy and applicability of a constructed multimodal description representation can be directly tested by applying to a route following task.

I collected a *spatial scene description corpus* where hand gestures and natural language are jointly used to describe spatial scenes. To focus on the multimodal and natural nature of scene descriptions, the data collection experiment was designed in a somewhat idealised setup, replacing real-life landmarks with simple geometric objects such as “circle” and “square”. This is to constrain the complexity of verbal descriptions and make the data more amenable to computational analysis (see below for details).

The corpus is composed of data from two experiments: the **spatial description experiment** (3.1.1) and the **scene description corpus** (3.1.2). While the former experiment focused on collecting intuitive multimodal descriptions, the latter one aimed to elicit multimodal data

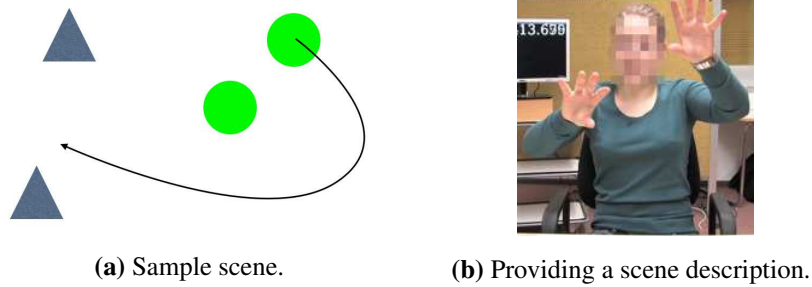


Figure 3.2: Providing a description in the Scene Description Experiment.

amenable to computational analysis and modelling.

In the rest of this section, I provide details on data collection and data analysis of each experiment.

3.1.1 The scene description experiment

Task design

To elicit intuitive scene descriptions, we designed a simple scene description task. Similar to previous works on spatial description tasks (Roy, 2002), instead of describing real-life landmarks, participants were asked to describe simple scenes composed of geometric objects and an arrow that indicates the movement of one of the objects, as shown in Figure 3.2a. While this setup retains the nature of multimodal descriptions, it effectively constrains the complexity on the language part, making it possible to model the descriptions using computational methods.

50 scenes were generated for the description task. Each scene is composed of 3 or 4 geometric objects (as shown in Figure 3.2a). The objects were in simple colours and were randomly placed in the scene. An arrow originates from one of the objects and points to another place in the scene. The shape of the arrow was not regular, hence it's difficult to describe the orientation and shape with natural language such as round or 90° to the left. Therefore, iconic gestures are needed to accurately describe the trajectory of the movement.

In route giving tasks, humans often give such descriptions from memory as the routes are often not in situated environment. To elicit natural behaviours in such descriptions, the scenes were shown on a computer screen for 10 seconds, then disappear before participants started to describe. To investigate whether and how humans combine gestures with language, participants were only asked to describe the objects, the spatial configuration and the movement indicated by the arrow. Gestures were not required. To encourage accurate descriptions, participants

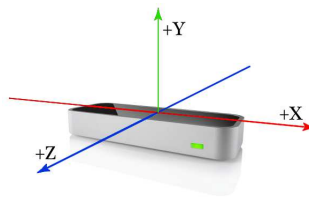


Figure 3.3: Leap sensor.

were told that another human will watch the video later and try to reproduce the described scenes. Describing the scenes as accurate as possible will make the reproduction task easier.

In the experiment, participants were seated in front of a table. A camera was placed in front of the participant to record audio and video data. A Leap sensor was placed on the table, right in front of the participant. The participant was told that the Leap sensor is a device to track their hand movements. The participant was allowed to play with the sensor and shown the effective tracking range of the Leap sensor. However, gestures were not mentioned in the instructions. On the right side of the table, there was a screen which displays scenes to be described. An experimenter sat next to the participant to control the display of the scenes. Each scene was shown for 10 seconds. After 10 seconds, the screen turned to black and the scene disappeared. Then participants started to describe as shown in Figure 3.2b.

Recording setup

I recorded the audio and video data 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.¹ The hand motion data was recorded using the MINT toolkit (Kousidis et al., 2013b).

Both video and hand motion data were timestamped to align hand motion data and natural language descriptions in the data processing procedure. The hand motion data was automatically timestamped by the MINT toolkit. To add timestamps in video recordings, we placed a monitor that displayed timestamps in front of the camera (behind the participant). Hence, the timestamps were recorded in the video.

Following hand features were recorded in the hand motion data (as provided by the Leap SDK):

- **FrameID:** integer, a unique ID assigned to this data frame.

¹www.leapmotion.com

- **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 with the observed data.
- **hand direction:** 3 dimension vector. The direction from the palm position toward the fingers.
- **hand sphere centre:** 3 dimension vector. The centre of a sphere fits to the curvature of this hand.
- **sphere radius:** float, the radius of a sphere fits to the curvature of this hand.
- **palm width:** float, the average width of the hand (not including fingers or thumb).
- **palm position:** 3 dimension vector, the centre position of the palm in millimetres from the Leap Motion Controller origin.
- **palm direction:** 3 dimension vector. The direction from the palm position toward the fingers.
- **palm velocity:** 3 dimension vector, the rate of change of the palm position in millimetres/second.
- **grab strength:** float, the strength of a grab hand pose as a value in the range [0..1], 0 when the hand is closed.
- **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].
- **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.
- **finger length:** float, the apparent length of a finger.
- **finger width:** float, the average width of a finger.
- **joint direction:** 3 dimension vector, the current pointing direction vector.

In the experiments, we found, some participants took it as a memory task and tried to describe with great details. For example, they used the distance (in *cm*) between objects and the image border to indicate object positions, which are common in natural scene descriptions. Some of them described with fewer details and rarely used gestures (or used gestures that are out of the effective tracking area), resulting inaccurate descriptions. We have also considered incorporating a confederate in the experiment. Although such face-to-face interactions would lead to spontaneous descriptions, that would also lead to clarification requests, immediate feedbacks and so on, which are not the focus of the study. Hence, we didn't incorporate confederates in our experiments.

In total, 14 participants took part in the experiment. All of them are native German speakers. On average, each of them finished 29 scene descriptions (SD = 9.60). 311.63 minutes of video and motion capture data were recorded. 179.51 minutes out of all the recorded video contain speech (57.6%). Next I describe the data processing and data statistics.

Data processing

A sampled scene 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] dem oberen gurrenden Kreis geht rechts neben dem anderen grünen Kreis 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] the upper green circle goes right next to the another green circle between the two triangles to the left of the arrow.

Transcription The audio recordings were manually transcribed by native German speakers. We then aligned the transcriptions with corresponding video recordings with a forced alignment approach (Baumann and Schlangen, 2012). The video and audio recordings were segmented into individual scene descriptions and annotated with a scene ID (e.g., *scene_1*). Each scene description was further segmented into object descriptions. We annotated each object description with an object ID. For instance, the object description in Example (2), *Hier_[deixis] ist ein graues Dreieck* was annotated as *object_1*, indicating that the described object is with ID 1 in the scene.

Gesture annotation Deictic gestures are conventionally divided into following gesture phases: pre-stroke, stroke, stroke hold and retraction (Kendon, 1980a). We manually annotated the stroke hold phase of each deictic gesture with ELAN², by watching the hand movements in the video recordings. Similar to the object description annotations, we annotated each deictic gesture with an object ID, such as *object_1*. The gestures and object descriptions with the same annotated IDs refer to the same objects in a scene description.

With the timestamps in the recorded videos and hand motion data, we aligned hand motion data with video recordings. Accordingly, the hand frames aligned with the stroke hold

²<https://tla.mpi.nl/tools/tla-tools/elan/>

annotations were labeled as *stroke hold* frames. The labels were used for training stroke hold detectors which will be elaborated in Chapter 5.

Data statistics

With the processed data, we conducted a preliminary data analysis in terms of: 1) Gesture space; 2) Episode length; 3) Scene matching error; 4) Re-reference precision; 5) Word number and description accuracy.

Varied gesture spaces I calculated the maximal gesture area that each participant's hands spanned during all their descriptions as their gesture space. As shown in Figure 3.4a, there are differences both within and between participants, making it a challenging task to detecting and interpreting deictic gestures with computational models.

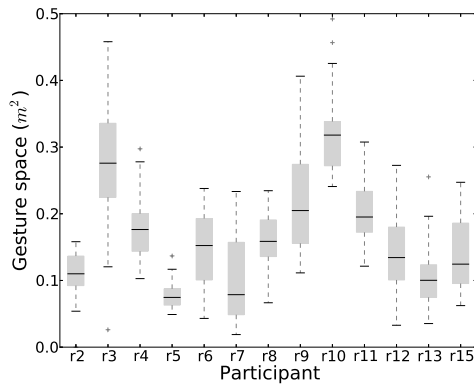
Dialogue time I calculated the duration of each scene description episode (in second). As shown in Figure 3.4b, similar to the size of gesture space, there are differences of episode length both within and between participants. An episode can span as long as more than 60 seconds, while in some cases as short as less than 10 seconds. Hence, we also analysed the relationship between the word number and corresponding gesture accuracy in each episode, as shown in Figure 3.4e. The results show that the more words spoken in an episode, the accurate the gestures are.

Matching error To investigate how accurately the deictic gestures represent the described the spatial layout of objects in the scenes, we adopted a shape matching method to compare the spatial layout of deictic gestures and corresponding objects. Figure 3.4f shows the histogram of the matching errors. The longer the verbal descriptions are, the accurate they references are.

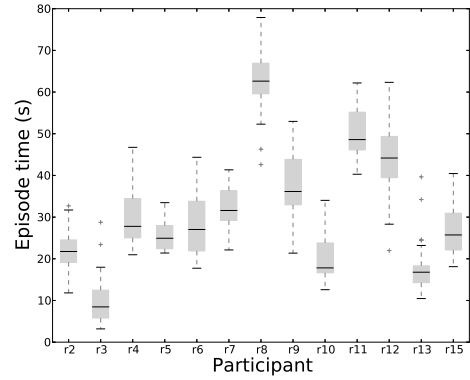
Re-referential accuracy I also investigated the re-reference accuracy in terms of the distance between re-reference spots and originally referred spots. In other words, how close when participants refer to a spot that has been referred to before? As shown in Figure 3.4c and 3.4d, among 185 re-reference points, 161 of them are with re-reference distance less than 150 mm, where gesture space is $900 * 671mm$.

3.1.2 The spatial description experiment

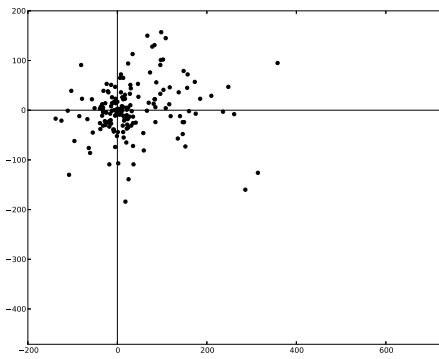
The data statistics in *scene description experiment* show that abstract deictics do accurately represent spatial layouts of objects in verbal descriptions. However, the largely variable gesture



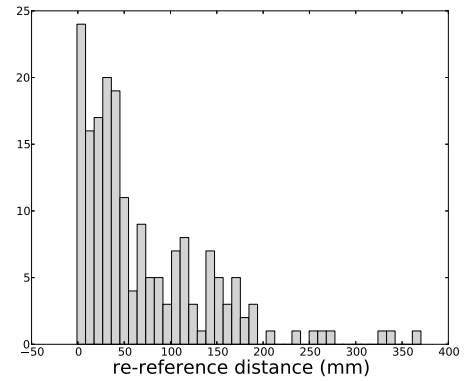
(a) Gesture space.



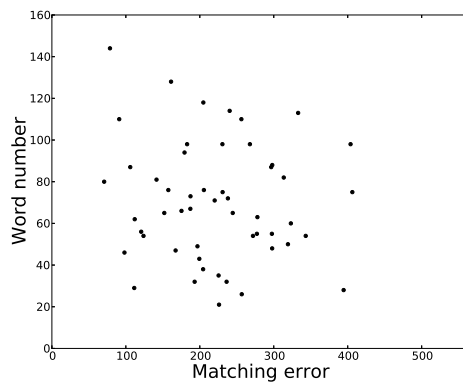
(b) Dialogue time.



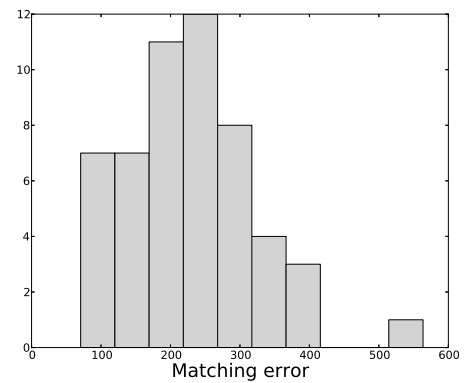
(c) Re-reference distance.



(d) Histogram of re-ref distance.



(e) Word number in each description and corresponding matching error.

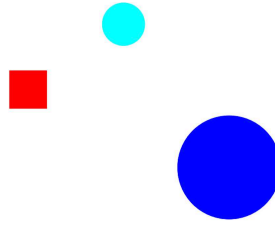


(f) Histogram of matching errors.

Figure 3.4: Data statistics of the Scene Description experiment.



(a) Providing a description.



(b) A sample scene of the task.

Figure 3.5: Providing a multimodal spatial scene description.

space makes it difficult to track all hand motions, which consequently makes it difficult to computationally modelling the interpretation task due to a lack of hand motion data. Therefore, we designed a *spatial description experiment* to collect multimodal description data which is more amenable to computational methods.

I designed the *spatial description experiment* with a similar task with the *scene description experiment*, but with human-computer oriented task instructions and simplified scenes, which will be elaborated below.

Task design

I designed a simple task (shown in Figure 3.5a) to elicit human-computer interaction oriented speech and gestures. I first generated 100 such scenes, each composed of two circles and one square. Figure 3.5b shows an example from the corpus. The size, shape and colour of each object were randomly selected when the scenes were generated. Object sizes are evenly distributed between 0.05 to 0.5 (as ratio to the image size). There are 6 colours and 2 shapes. Each of them had equal chance to be assigned to an object. The object positions were randomly generated and adjusted until none of them overlap with each other.

In the experiments, the scenes were shown in the same order to all participants. For each description episode, a scene was displayed on a computer screen. Participants were told that they were talking to an automated system. Encoding object properties like colour, shape, size, positions of objects, and demonstrate the relative positions with deictic gestures will make it easier for the computer to understand the descriptions. After each description, there was a score on the screen, ostensibly reflecting how well the system understands the description. In reality, the score was given by a confederate who had the instruction to rate the description based on the number of mentioned attribute types. Participants were asked to describe as accurately as possible and try to get a high score so that the computer (in reality, the Wizard) can understand

the description.

Recording setup

In the experiment, hand motion was also tracked by a Leap sensor³ as in the scene description experiment. However, in this experiment, a new Leap SDK (SDK v2.3.1) was used. It not only gives *hand number* in the tracking area, but also provides the **hand type** such as “left hand ” or “right hand”. With the new SDK, it’s possible to label all hand motion frames of each hand over time. The same as the scene description experiment, the hand motion data was recorded with MINT tools (Kousidis et al., 2013b). Audio and Video were recorded by a camera. Timestamps were recorded in videos and hand motion data.

After introducing the task, participants were seated in front of a desk. A monitor was on the right of the desk to show the scenes to be described. A Leap sensor was on the desk in front of participants for tracking hand motion. Due to the small tracing area of the sensor, we set a monitor in front of participants to display hand movements, so that they can see whether their hands were tracked while they were gesturing. This helps to track all hand movements. None of the participants reported unnatural gestures due to limited gesture space. Before the experiment, participants had several minutes to play with the sensor, so that they know the boundaries of the tracking area. When they got familiar with the sensor, the experimenter gave instructions and demonstrated how to describe a scene with speech and gestures. The experiment started after the participant confirmed that she/he understood the task. Then the monitor on the right showed a scene. After looking at the scene for a few seconds (they could look at it as long as they want until they are ready for descriptions), participants started to describe. Shortly after the description ended (around 1 second), a score was shown on the screen for 10 seconds, then the wizard advanced to the next scene. There was no time limit for each scene description.

Data preprocessing

Example (3) shows a sample multimodal description from the corpus:

- (3) Hier_[deixis] ist ein kleines Quadrat, in rot, hier_[deixis] ist ein hellblauer kleiner Kreis und hier_[deixis] ist ein blauer grosser Kreis.
Here_[deixis] its a small square, red, here_[deixis] is a light blue small circle and here_[deixis] is a blue big circle.

The recordings were manually transcribed by native speakers. The transcriptions were aligned

³www.leapmotion.com

with the videos via an automatic forced alignment approach. Utterances for each object were manually annotated with the referred objects. For example, “*hier ist ein kleines Quadrat*” (here is a small square) in the above example might be annotated as *object1* in the scene image.

The deictic gestures were manually annotated based on hand movements in the videos with ELAN, a software for annotation⁴. They were also annotated with referred objects in the same way as the utterance annotation. With the recorded timestamps, hand motion data were aligned with videos. Aligned hand motion data frames were labeled as *stroke hold* frames (hand stays at the targeted position Kendon (1980a)) or *non-stroke hold* frames (hands in movements or not refer to any target object).

Data statistics

In total, 15 participants (students from Bielefeld University) took part in the experiment. 14 of them are native German speakers. One participant only described verbally. We excluded the non-native speaker and the one who didn’t gesture. As a result, 311.75 minutes recordings with audio and video were collected. 830 scenes were described. Below, I give data statistics in terms of verbal descriptions, gestural behaviours, temporal and semantic relations between deictics and accompanied verbal descriptions.

From the recorded data, we found that, when describing an object, participants gestured either with one hand to denote the location of the object or with two hands to denote the relative position to another object. In both cases, spatial layout of objects are encoded in gestures. The frame rate of hand motion data was around 100 frames per second as recorded by the Leap sensor. Hence, the recorded data are sufficient for real-time and incremental processing.

Variability of verbal descriptions Although participants were instructed to mention colour, shape, size and relative positions of the objects, they were allowed to formulate the descriptions in their own way. In other words, the verbal descriptions are, within these constraints, natural descriptions, and they do indeed vary in how the information was formulated and linearised:

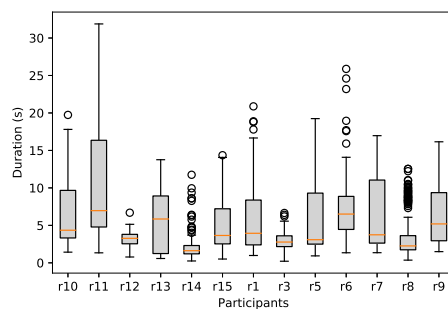
1. Varied object attribute descriptions. The vocabulary size of the corpus is 294, with 16891 tokens. On average, the vocabulary size for each attribute description is over 20. For instance, participants used different words to describe the same colour and shape. Purple and cyan objects were also described as *lilac* and *light blue*. Circles were also referred as *balls*. An object at bottom left was also described as *a bit lower* to another object.
2. Flexible information sequence. While participants often describe positions (i.e., bottom left) first, followed by size, colour and shape information (position, size, colour, shape).

⁴<https://tla.mpi.nl/tools/tla-tools/elan/>

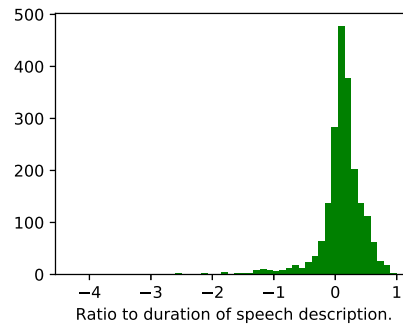
Some of the descriptions started with colour, shape, with size and position information described at last (colour, shape, size, position) or in other sequences. On the other hand, some words occur frequently at the end of object descriptions. For example, the shape of objects (i.e., circle) usually encloses an object description. Hence, these words are informative features for the object description segmentation task, which will be described in more details in Chapter 5.

3. Flexible gesture/speech compositions. Some participants supplement deictic gestures with position words like “bottom left”, they also describe positions with deictic word “here”. In the former case, the gesture and speech supplement each other, while in the latter case, they complement each other. Alternatively, position information was simply ignored in verbal descriptions.

Spatial descriptions According to the way position information is expressed, the collected spatial descriptions fall into two categories: absolute position or relative position. For example, the position of an object can be described according to its absolute position in the scene such as *bottom*, alternatively, it can also be described relatively to another object in the scene, such as *a bit lower to the red circle*. Presumably, the latter expression should occur only when describing the second or third object. Moreover, the description “a bit lower” is ambiguous as the exact spatial layout between two objects are not encoded in the speech. In this case, a deictic gesture can effectively disambiguate the layout by demonstrating the spatial relation with the positions of deictic gestures.



(a) Duration of deictic gestures.



(b) Temporal relations of speech and deictics.

Statistics of deictic gestures There are 2138 deictic gestures out of 2490 object descriptions in the corpus. Figure 3.6a shows the average gesture durations of each participant. The overall average duration of a deictic gesture is 5.22 seconds, with a maximum duration 31.9 seconds

and a minimum duration 0.21 seconds. As shown in the Figure, the durations vary both within and across participants. Four of the participant gestured for less than 5 seconds, while other participants gestured longer.

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 3.6b. 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), as we will show in Chapter 5.

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.

3.2 Multimodal object description corpus

In this section, I introduce the *Multimodal Object Description Corpus*, a dataset that pairs objects in real-life photographs with natural language description and hand-drawn sketches, as shown in Figure 3.7. While the verbal description “facing right, trunk coiled towards mouth” describes the overall orientation of the elephant and the shape of the trunk symbolically, the hand drawn-sketches visually signifies the exact shape of the elephant’s nose. In other words, the sketch enriches the verbal description with visual information.



(a) Hand-drawn sketch collected in (Sangkloy et al., 2016).



(b) Photograph of an elephant.

Figure 3.7: Example of a multimodal description: *facing right, trunk coiled toward mouth.*

The multimodal object description corpus provides symbolic and iconic descriptions of natural photographs with words and sketches respectively. The corpus is composed of natural photographs of objects, sketches and verbal descriptions of each object in the photographs. The photographs and sketches are from an existing corpus – the Sketchy Dataset⁵, which was originally introduced in Sangkloy et al. (2016). I started from the Sketchy Dataset and augmented 86% sketches in the dataset with verbal descriptions (see Section 3.2.2 for details). In the rest of this section, I first describe the selection of photographs, the Sketchy Dataset and the crowdsourcing experiment conducted to collect the verbal descriptions.

3.2.1 The Sketchy dataset

The Sketchy Dataset (Sangkloy et al., 2016) includes 12500 unique photographs of real world objects that expand over 125 categories, and 75471 sketches paired with the photographs.

Photographs The photographs in the Sketchy Dataset were selected from ImageNet (Russakovsky et al., 2015). Each photograph contains exactly one object with an annotated bounding box. These photographs were selected according to following criteria: (1) exhaustive, the categories should cover a large number of common objects; (2) recognisable, each category should have recognisable sketch representations, so that the sketches can be distinguished from sketches in other categories; (3) specific, each object should be sketch-able so that the sketches of the objects are not uninformative;

To increase visual diversity and the number of “sketch-able” photographs in each category, some categories in the ImageNet dataset were combined into same categories. The categories with fewer sketch-able images than 100 were excluded from the corpus. Each photograph is

⁵<http://sketchy.eye.gatech.edu/>

annotated with a subjective “sketch-ability” score which ranges from 1 (very easy to sketch) to 5 (very difficult to score). The 100 images in each category are with a targeted sketch-ability distribution of 40 very easy, 30 easy, 20 average, 10 hard and 0 very hard.

For more detailed descriptions of the image selection procedure, please refer to the original paper Sangkloy et al. (2016).

Sketches The sketches were collected by Sangkloy et al. (2016) via Amazon Mechanical Turk⁶ (AMT). Workers were shown a random photograph from the photograph database, and instructed to sketch the named object with a similar pose on a canvas. They were instructed to only sketch the target object and avoid other areas of the photographs. In each episode, workers click a button to display a photograph. The photograph was shown on the screen for 2 seconds, then hidden before participants start to sketch. To make sure that the sketches are drawn from memory, a visual noise mask was displayed after the photograph disappeared, so that the low level visual representations in the working memory were destroyed (Grill-Spector and Kanwisher, 2005; Nieuwenstein and Wyble, 2014). This ensures that the sketches are realistic and diverse. As a result, the sketches implicitly encode salient visual information of objects, but are different from boundary annotations (Lin et al., 2014; Xiao et al., 2016). Workers can view the photograph as many times as they want by clicking a button, however, after each viewing, the canvas was cleared.

Each photograph was paired with several sketches from different workers to capture diversities. The sketches were stored as SVG files, which include high resolution timing details. For each stroke, the start and end times, along with fine-grained timing along a stroke are included in the SVG files. The recorded information of stroke length and order enables us to reconstruct the sketches, render sketch images with different percentage of strokes and investigate the contribution of reduced sketches (see Chapter 6 for more details).

3.2.2 Augmenting sketches with verbal descriptions

To form sketch-verbal description ensembles for each photograph, we augmented the Sketchy Dataset with verbal descriptions. Each photograph was paired with a verbal description. The description can form different ensembles with different sketches for each target photograph.

⁶<https://www.mturk.com>



Figure 3.8: Discriminative description of the left-most photograph provided by crowdworker: *facing right, trunk coiled toward mouth*.

Data collection

I used the Crowdfower service⁷ to collect object descriptions from English speakers. Since the category of each photograph has been pre-specified, photographs from different categories can be distinguished by the category name. Hence, we designed this augmenting experiment as a within category image selection task. This helps to collect object attributes (for example, colour, orientation, shape, etc.) that distinguish an image from images in the same category. While some attributes such as colour are impossible to be encoded in sketches, object orientations (e.g., facing right) can be both encoded in verbal descriptions and sketches. In the former case, the verbal description complements sketches with information in symbolic mode; in the latter case, the verbal description supplements the sketches.

In the experiment, workers were shown 7 images from the same category for each photograph description task. As shown in Figure 3.8, the target image was shown in a larger size. The distractor images were shown in rows of 3 side-to-side with the target photograph. In this example, the worker listed two attributes of the elephant in the target photos: *facing right* and *trunk coiled toward mouth*. Note that the distractor image at row 2, column 3 also fits with description. Thus the description is even confusing for humans. Only when coupling the description with the sketch of the photograph in Figure 3.7, it's possible to resolve the target photograph.

At the beginning of the description task, workers were required to read an instruction of how to perform the task. They were instructed to list all attributes that might help a human to distinguish the image from distractor images. Attributes such as colour, shape, size and

⁷<https://www.crowdfower.com>

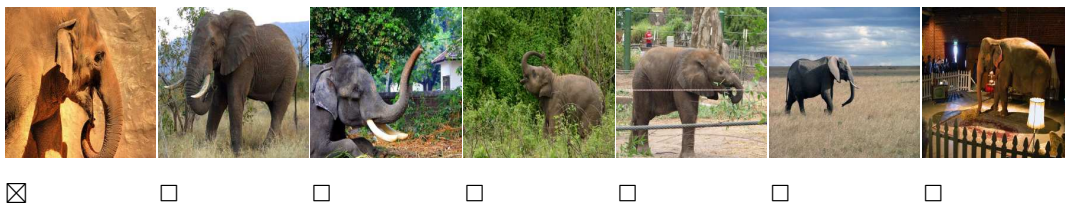
orientation were suggested, but workers were encouraged to list any attribute that might help. Workers were asked to separate each attribute with a “,”. They were told that no need to type the object category name of the target object as all the objects are of the same type, thus the descriptions are only about object attributes. In following sections of the dissertation, I will refer to these words as **attribute words** (+ATT), to distinguish them from object category names such as “*elephant*”, which are referred as **category words** (+CAT).

Since this is a subjective task, no words or attributes were listed, from which workers could select. To ensure the description quality, we set a minimum vocabulary size of each description of 5. In other words, workers should use at least 5 different words in each description to describe the target object.

Data validation To get an idea of whether the collected verbal descriptions are informative to distinguish the target image from distractor images, we conducted a small scale data validation experiment.

The target images and corresponding distractor images were shown to workers on the Crowdfunder platform. The images were randomly arranged and shown side-by-side. An object description was shown on the bottom of the images. Workers were instructed to select the image that fits best with the description by clicking the check box under the image.

100 descriptions were randomly selected and shown to workers. Among all descriptions, 70% of them were correctly selected with the given verbal description by workers. It shows that, even for humans, the verbal descriptions can be confusing.



Please select the image that fits best with following description:

facing right, trunk coiled toward mouth.

Figure 3.9: Example of data validation test.

3.2.3 Data statistics

In total, 10,805 object descriptions were collected. After running a spell checker to correct typos, there are 100,620 tokens in all descriptions. The vocabulary size is 4,982. The ratio

between types and tokens hence is 0.5. On average, each photo was annotated with 3 attributes. Each of the attributes on average spans 4.6 words.

I augmented 10, 805 unique photographs with verbal descriptions. 64 905 sketches were paired with these photographs in the Sketchy dataset. Consequently, there are 64 905 sketch-verbal description ensembles.

3.3 The SAGA corpus

The SAGA corpus, originally introduced in Lücking et al. (2010), contains dialogues between route givers and route followers. It also provides fine-grained annotations for speech and gestures in the dialogues. Based on the annotations, we conducted our experiment in Chapter 7.

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 video recordings; 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 dataset, each item is a tuple of 4 elements: (*words, gesture features, speech semantic categories, gesture semantic categories*).

There are 5 gesture semantic category labels in the annotations: *shape, size, direction, relative position, amount*; the speech semantic labels consist of the above labels 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. From these the multimodal 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 in this dissertation, I focus on speech and gesture semantics, I only 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:

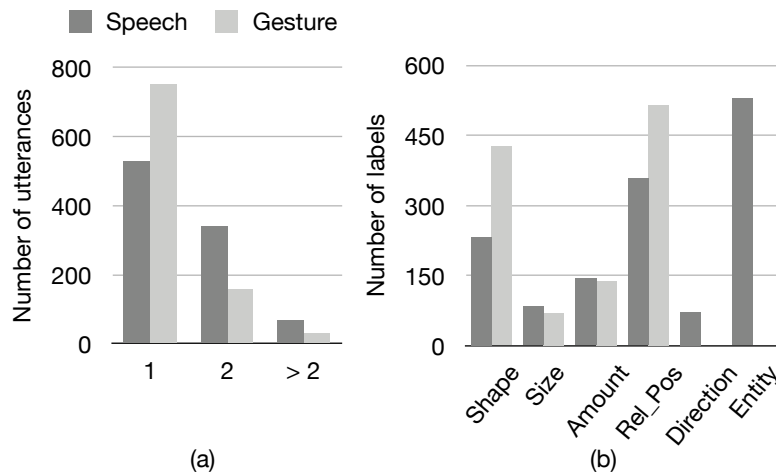


Figure 3.10: (a) Histogram of semantic labels per utterance/gesture. (b) Histogram of semantic labels. (Rel_Pos indicates relative position.)

{*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 3.10 (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 3.10 (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 in the SAGA corpus, we used the manual annotations to represent gestures. For instance, the gesture in Figure 7.1 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.

These annotated tokens were treated 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.

3.4 Summary

In this chapter, I first described the data collection, data processing and data statistics of two corpora: the *Spatial Scene Description Corpus* and the *Multimodal Object Description Corpus*. While the first corpus provides multimodal descriptions composed of natural language and deictics, the latter one provides object descriptions composed of natural language and iconic elements (i.e., hand-drawn sketches). I also briefly introduced the SAGA corpus which provides multimodal dialogues composed of natural language and hand gestures. In the following chapters, I will elaborate the computational models that are built based on these datasets.

4

A system of understanding multimodal spatial descriptions

In this chapter, I present a system of modelling the interpretation of multimodal spatial scene descriptions composed of abstract deictics and verbal utterances. The system is supposed to take speech and hand motions of spatial descriptions as input, understand speech and hand gesture information, represent the knowledge in a format that can be perceptually grounded to scene images, then apply the knowledge to perform a scene retrieval task. This chapter focuses on the exploration and evaluation of 3 knowledge representation variants: **a)** verbatim representation; **b)** representation with pre-specified symbols; **c)** representation with automatically learned symbols. I will first provide an overview of the system and introduce individual system components, then describe the 3 representation variants. This is followed with an evaluation experiment and discussions of the evaluation results.

4.1 Modelling the interpretation of multimodal spatial descriptions

In the previous chapter, I introduced the Multimodal Spatial Description corpus. Aiming to computationally modelling the interpretation of multimodal spatial descriptions, this chapter describes a multimodal system that can interpret and apply multimodal spatial descriptions.

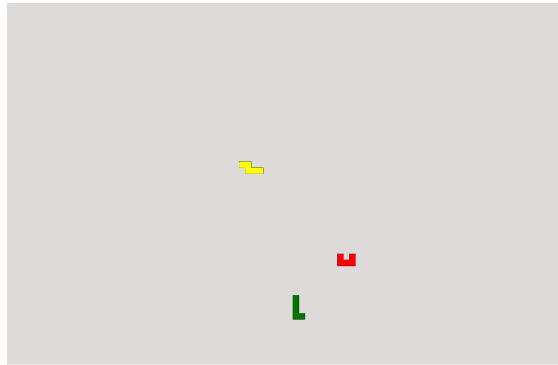


Figure 4.1: Scene example.

As such as system contains several individual modules, in this chapter, I particularly focus on introducing and discussing 3 representation variants of the multimodal content, with a constrained setup.

Figure 4.1 shows an example scene that was used in this task. An example scene description of the task is shown as follows:

- (1) *Top left _[deixis] is a yellow Z and bottom, in the middle _[deixis] a green L, and bottom right ...*

With the multimodal scene descriptions such as illustrated by example (1) from a human speaker, a human listener typically forms a mental image of the described scene, where visual attributes of the referents (i.e., the objects) are introduced via natural language; positions of the referents are demonstrated via deictic gestures; the spatial configuration of the objects are contrasted both via gesture positions and natural language (*slightly above*), represent the knowledge of the image in his mind, then later apply the knowledge to recognise the described scene in real world.

I'm interested in modelling the listener's task with a computer system. Such a modelling task forms a good test case of constructing representations of multimodal descriptions. Given a verbal and deictic scene description as in Example (1) in a real-time manner, the system constructs a representation of the multimodal description. After building the scene representation, the system applies the representation to candidate scene images to retrieve the target scene, which best conforms to the description. The task hence requires to **1**) segment perceived speech and gestures into individual object descriptions; **2**) construct the representation based on object descriptions; **3**) apply the representation to a visually perceived context.

$$(2) \quad \begin{array}{|l} \hline o_1, g_1, o_2, g_2 \\ \hline o_1: \text{transl}(\text{red circle}) \\ g_1 : (x_1, y_1) \\ \quad \text{pos}(o_1, \phi(g_1)) \\ \text{slightly_above}(o_1, o_2) \\ \\ o_2: \text{transl}(\text{blue L}) \\ g_2 : (x_2, y_2) \\ \quad \text{pos}(o_2, \phi(g_2)) \\ \hline \end{array}$$

Inspired by the discourse representation tree model (Lascarides and Stone, 2009), the referents of objects were represented with a set of symbols. In (2), $\text{transl}()$ indicates a function that translates utterances into logical forms. The co-verbal deictic gestures, which specify positions of referents are represented as g_i , connected to corresponding object referents. In current task, we assume that deictic gestures only encode positions information on a 2-D plane which can be represented as (x_i, y_i) . $\text{pos}(o_i, \phi(g_i))$ is a function which transforms a raw hand position to its logical form.

In particular, we explore three variants of representations of the verbal content, ranging from verbatim representations, representations with pre-defined property symbols to representations with a set of automatically learned symbols, which will be described in Section 4.2.2.

The performance of the retrieval task gives a practical measure for the quality of the representation. If the representation does indeed capture the relevant content, it should form the basis for identifying which scene was described from distractor scenes. Therefore, we constructed a system with a pipeline for processing and representing such multimodal descriptions, applied and evaluated the representations with a scene retrieval task.

In the rest of this chapter, I first gives an overview of the system, then describe the three representation variants. This is followed with an experiment that evaluates the system performance under different setups.

4.2 System overview

Figure 4.2 shows an overview of the system framework. The system is composed of two processing pipelines: a language processing pipeline and a gesture processing pipeline.

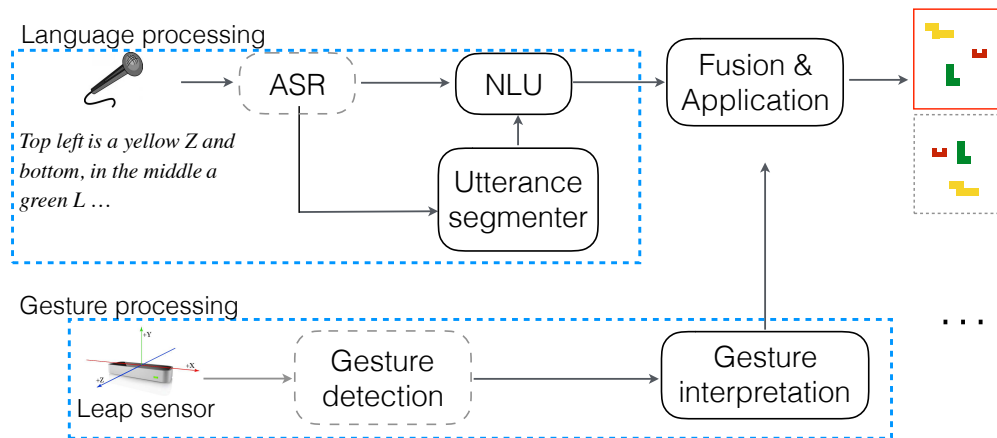


Figure 4.2: Overview of the system framework. (Modules in grey boxes are not implemented in this chapter, but simulated. See the Experiment section for details.)

Language processing pipeline In an ideal setup of the system, the speech is processed by an ASR which produces output word-by-word. The output then is fed into a segmentation module that determines the boundary of each object description. In other words, it decides when a new object is introduced in the discourse. Once a segmentation signal is received, the signal is sent to the NLU module which initialises its model for the incoming description (i.e., create a new NLU model for each object description).

Gesture processing pipeline In parallel with the language processing pipeline, a motion capture sensor records hand motion data and sends the data to a deictic gesture detector (in the work of this Chapter, we simulated the detector with object positions in scenes, see 4.5 for details). This detector sends a signal to the Representation module when a deictic gesture is detected. The representation module represents all the verbal descriptions and deictic gestures as a scene description until a new scene description starts, then the module creates a new scene representation for the incoming scene description.

Multimodal fusion & application After the full description of a scene has been perceived, the representation is used to make a decision in the scene retrieval task. The candidate scenes are given by a computer vision module, which recognises the objects in the scenes and computes a feature vector for each, containing information about the colour of the object, the number of edges, its skewness, position, etc; i.e., crucially, the object is not represented by a col-

lection of symbolic property labels, but by real-valued features. The application module takes this representation for each candidate scene as input, and computes a score how well the stored representation of the description content matches the candidate scene. For this, it makes use of the perceptually-grounded nature of the symbols used in the content representation, which connects these with the object feature vectors. The scene with the highest score finally is chosen as the one that is retrieved.

We now describe some these processing steps in details.

4.2.1 Utterance segmentation

The task of the utterance segmentation module is to identify the boundaries between object descriptions. That is, the utterance segmenter segments an object description into individual utterances of each object description and informs the NLU module when a new object description starts. For example, a description such as “top left is a red L a bit left of it is a yellow T” is expected to be segmented into two utterances: “top left is a red L” and “a bit left of it is a yellow T”. In this work of exploring representation methods, we adopted a dataset with key words that signify utterance segmentations (see Section 4.5 for details), hence the utterance segmenter is a simple rule-based approach. Chapter 5 presents a learned segmenter model.

4.2.2 Representing scene descriptions

Within the above basic structure of scene descriptions, we explored three variants of scene representations:

| | DESCRIPTION | REPRESENTATION | APPLICATION | VISUAL INPUT |
|---|------------------|-----------------|----------------------|--------------------------------------|
| A | speech + gesture | word stems | word classifiers | raw visual object and scene features |
| B | | property labels | property classifiers | |
| C | | cluster labels | cluster classifiers | |

Table 4.1: Overview of representation variants A-C.

- In **Variant A**, we translate the referring expressions simply into a sequence of lemmata. This would lead to a representation of “red circle” as *red, circle*.
- In **Variant B**, the translation proceeds by specifying a semantic frame, but here by way of more practically oriented approaches to spoken language understanding for object descriptions, leading to, for example for “red circle”. This presupposes availability of a process that can do such a mapping; e.g., a lexicon links lexical items and such frame

elements, and a pre-specified repertoire of attributes and values for them. In this work, there are 5 pre-defined properties: *colour*, *shape*, *orientation*, *horizontal position* and *vertical position*.

- In **Variation C**, finally, we map the referring expression into a sequence of symbols (similar to Variation B) where however the repertoire of these symbols comes from an automatic learning process, and thus does not necessarily correspond to pre-theoretic notions of the meaning of such attributes.

For all the three variants, the represented symbols in the representations are perceptually grounded to objects in real images. In other words, the representations can be applied to a task and evaluated with the applicability.

Table 4.1 gives an over view of the three representation variants and corresponding applications. In following section, I describe the application which we used to apply and evaluate our representation methods.

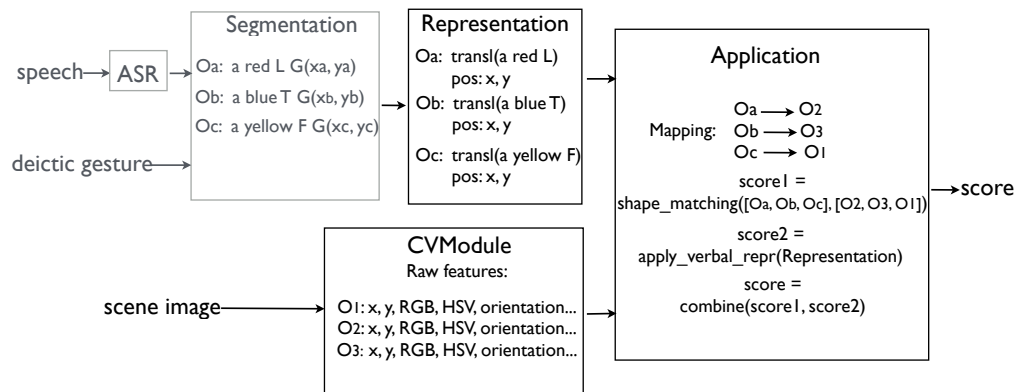


Figure 4.3: Processing pipeline

4.2.3 Applying gestural information

As described above, the discourse representation includes position information of objects indicated by the deictic gestures. To make use of this information to reconstruct the spatial layout of the described scene and subsequently apply the representations to retrieve the target scene, the first step is to compute for each scene the likelihood that it, with the position that objects are in, gave rise to the observed (and represented) gesture positions.

This is not as trivial as it may sound, as the gesture positions are represented in a coordinate system given by the motion capture system (i.e., the Leap sensor coordinate system), whereas the object positions are relative to the image coordinate system. Moreover, the gestures may

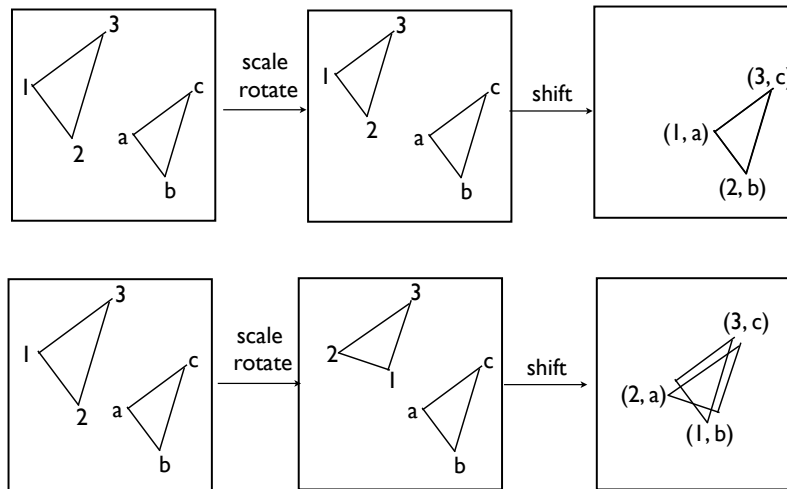


Figure 4.4: Example of a good mapping (top) and bad mapping (bottom), numbered IDs represent the perceived objects, the letter IDs represent the described objects.

have been performed sloppily, which lead to noisy position information and inaccurate spatial layout. Finally, on a more technical level, the labels that the segmentation module assigned to the parts of the description (O_a etc. in Figure) don't immediately map to those given to the objects recognised by the computer vision module (O_1 etc.).

To address the question: *which description object to compare with which computer vision object*, we simply perform exhaustive mappings - try all permutations of mappings. For each mapping a score is then computed for how well the gestured configuration under a given mapping can be transformed into the scene configuration.

The transformation procedure is illustrated in Figure 4.4. First, we project the gesture positions into the same coordinate system as the scene configuration, then it's scaled, rotated and shifted to be as congruent with the scene configuration as possible. In Figure 4.4, where the top target mapping between description object IDs and scene object IDs is sensible, the operation leads to a good fit, the bottom mapping is not as good.

Technically, the mapping works as follows. The positions of the three objects in the description and in the visual scene can be represented as matrices S_d and S_v of the form:

$$S = \begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \\ x_3 & y_3 \end{pmatrix} \quad (4.1)$$

with a set of parameters p

$$p = [\theta, t_x, t_y, s] \quad (4.2)$$

where θ is the rotating angle; $t_x(t_y)$ stands for the shift value on the $x(y)$ axis; s is the scaling parameter. We scale, rotate and shift matrix S_v to derive a transformed matrix S_t :

$$S_t(x, y) = \begin{pmatrix} t_x \\ t_y \end{pmatrix} + s \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \quad (4.3)$$

By minimizing the cost function:

$$E = \min \| S_t - S_d \| \quad (4.4)$$

We compute the optimal value for p . The distance between the resulting optimal S_t and S_d gives a metric for the goodness of the mapping, which is the input for a likelihood model that runs into a probability.

4.3 Learning knowledge from prior experience

We assume that our system brings with it knowledge from previous experience with object descriptions. This knowledge is used (at least in some variants) for the task of mapping to logical forms, and in all variants for the perceptual grounding of the symbols in the logical form. In what follows, I first briefly describe the corpus of interactions from which this prior knowledge is distilled.

4.3.1 The TAKE corpus

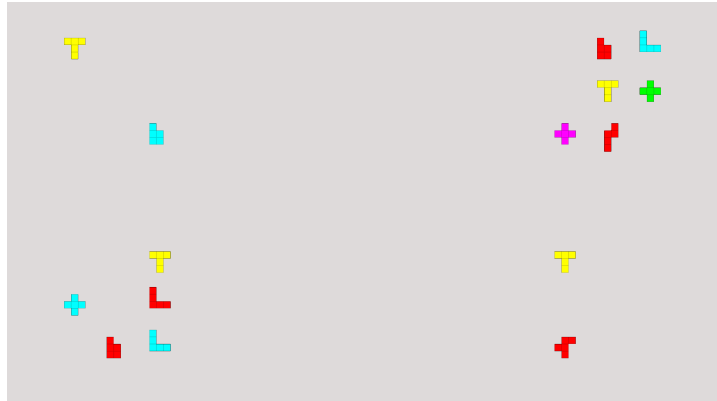


Figure 4.5: Example TAKE scene used for training.

As a source for learning prior knowledge, we use the TAKE corpus (Kousidis et al., 2013a). In a Wizard-of-Oz study, participants were presented with a scene of pentomino pieces (as shown in Figure 4.5) on a computer screen and asked to identify one piece to a “computer system” by describing and pointing to the piece. The utterances, arm movements, pentomino scene states and gaze information were recorded as described in (Kousidis et al., 2012). In total, 1214 episodes were recorded from 8 participants (students from Bielefeld University; native German speakers). The corpus was further processed to include raw visual features of each piece tile such as colour, shape, HSV, RGB values and so on. The computer vision module in our system processes the objects in the scenes in the same manner. It also includes symbolic properties (e.g., *green*, *X (a shape)*) for the intended referent, and the utterance that the participant used to refer to the target referent.

4.3.2 Learning mappings to logical forms

As described above, the difference between the 3 variants of representations lies in how they realise the *transl()* function to encode multimodal descriptions to representations in a computer system. In all variants, there is a preprocessing step that normalises word forms by *stemming* them using the NLTK (Loper and Bird, 2002). This will map all words to its stems. For example, we mapped all of *grün*, *grüne*, *grüner*, *grünes* into *grun* (*green*). This effectively reduces the vocabulary size that needs to be mapped.

Variant A

In Variant A, we stem each word, then treat each word stem as a logical form. In other words, an object description is translated into the sequence of its stemmed words. Hence, this variant contains the largest amount of logical forms.

Variant B

In Variant B, similar to the model presented in Kennington et al. (2013), we define a set of symbolic property labels. These labels are used to represent each described entity. Then we map each description from words to these symbolic labels, based on co-occurrence in the training data. For instance, if a word *grün* (*green*) frequently occurs when the described referent has the property *green*, we strength the link between the word *grün* with the symbolic label *green*.

Given a word (word stem in this case), the model returns a probability distribution over all properties; we average over the contribution of all words, and chose the most likely property as the representation for the description. Note that this variant does not require a pre-specified

lexicon linking words to object properties (e.g., *green*, *red* etc. in total 7 colour and 12 shape properties).

Variant C

In Variant C, we overcome the limitations of Variant A and B by automatically learning a set of symbolic labels from the data, rather than pre-defining symbolic labels as in Variant B or using word stems as in Variant A. As will be described below, for Variant A we learn for each word stem a classifier that links it to perceptual input. These classifiers themselves can be represented as vectors (the regression weights of the logistic regression). Using the intuition that words with similar meaning should give rise to similarity behaving behaviours (e.g., the classifier to “light green” should respond similarly - but not identically - to the classifier for “green”). We ran a clustering algorithm (K-means clustering using the Scikit-learn package¹) on the set of classifier vectors. The resulting clusters, through their centroids, can then themselves be turned again into classifiers. This effectively reduces the number of classifiers that need to be kept in a computer system, just as in Variant B, the set of properties is smaller than the set of words that are mapped into it. In contrast to Variant B, here the clusters are chose based on the data, rather than on prior assumptions.

With the above described method, an object description is represented as a sequence of the labels of those clusters that the words in the description map into, such as $\{c_1, c_3, c_{20}, c_1\}$.

After mapping words to logical forms, I now describe how we ground these logical forms to visual features of objects.

4.3.3 Learning perceptual groundings

Variant A

For Variant A, we learned grounded word stem meanings in a similar way as done in (Kennington et al., 2015; Schlangen et al., 2016). For each word w occurring in the TAKE corpus of referring expressions, we train a binary logistic regression classifier (see Equation 4.5 below, where w if the learned weight vector and σ is the logistic function) that takes a visual feature representation of a candidate object x and return a probability p_w for this object being a good fit to the word. We present the object that the utterance referred to as a positive training example for a good fit, and objects that it didn’t refer to as a negative example. (see (Kennington et al., 2015; Schlangen et al., 2016) for more details for the grounding method).

¹<http://scikit-learn.org/>

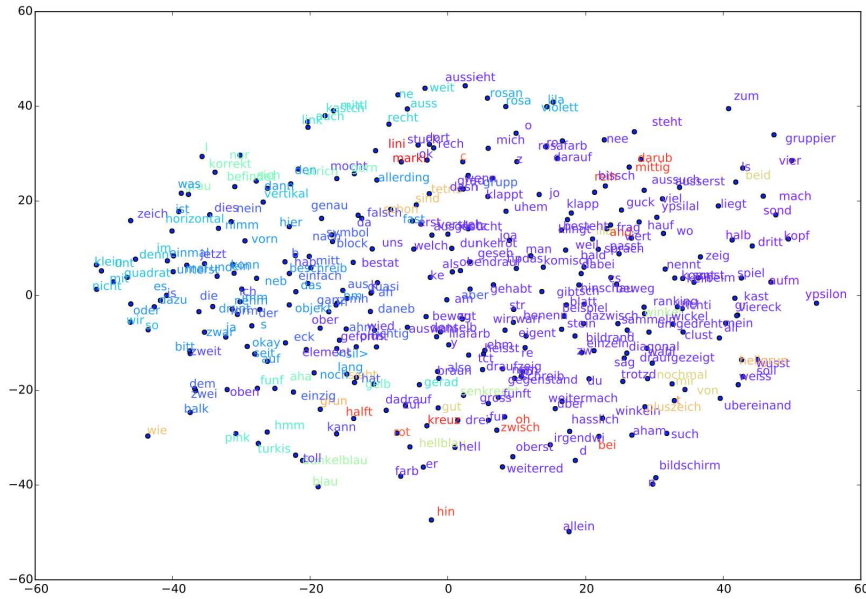


Figure 4.6: Clusters of words according to the co-efficient values of corresponding word classifiers. See 4.3.2 for detailed descriptions.

$$p_w(\mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b) \quad (4.5)$$

As aforementioned, each classifier is fully specified by its coefficients \mathbf{w} and b .

Variant B

The first step in Variant B was to use the words in the object description as evidence for how to fill the semantic frame, with the frame elements *colour* and *shape*. For the possible values of these elements (e.g., *green*), we trained the same type of logistic regression classifier, again using cases where the property was presented for a given object as positive example, and, as negative examples, those where it wasn't. This then gives a perceptual grounding for the property *green* (whereas in Variant A we trained one for the word *green*). In a way, this variant raises the question of where the ontology of properties comes from; if this part is a model of language acquisition, the claim would be that there is a set of innate labels which just need to be instantiated.

Variant C

As described above, Variant C is built on Variant A, reducing the required set of classifiers through clustering. In the experiment described below, we set the number of clusters to 26, an experimentally determined optimum. The clustering resulted for example (as shown in Figure 4.6) in one cluster grouping together “violett” and “lila” (*violet and purple*), or another one clustering group “türkis, blau, dunkelblau” (*turquoise, blue, dark blue*), but also clusters that are less readily interpreted such as “nochmal, rosa, hmm” (*again, pink, erm*). What is important to note here is that in any case, the reduction in the range of what words can map into in their semantic representation is as strong as with B, but emerges from the data.




| Description | Representation | Mapping | Perception/Scene |
|------------------------|---|---|---|
| “here red T” (1, 3) | A → “here red T” | $\text{avg}(C_{\text{here}}(x_1) + C_{\text{red}}(x_1) + C_{\text{T}}(x_1)) * P((1,3) (1,3)) = 0.4$ $\text{avg}(C_{\text{here}}(x_2) + C_{\text{red}}(x_2) + C_{\text{T}}(x_2)) * P((1,3) (1,3)) = 0.6$ $\text{avg}(C_{\text{here}}(x_3) + C_{\text{red}}(x_3) + C_{\text{T}}(x_3)) * P((3,1) (1,3)) = 0.3$ | X_1 (1,3)  |
| | B → color: red shape: T | $\text{avg}(C_{\text{red}}(x_1) + C_{\text{T}}(x_1)) * P((1,3) (1,3)) = 0.4$ $\text{avg}(C_{\text{red}}(x_2) + C_{\text{T}}(x_2)) * P((1,3) (1,3)) = 0.62$ $\text{avg}(C_{\text{red}}(x_3) + C_{\text{T}}(x_3)) * P((3,1) (1,3)) = 0.3$ | X_2 (1,3)  |
| | C → cluster ₁ cluster ₅ cluster ₃ | $\text{avg}(\text{cluster}_1(x_1) + \text{cluster}_5(x_1) + \text{cluster}_3(x_1)) * P((1,3) (1,3)) = 0.4$ $\text{avg}(\text{cluster}_1(x_2) + \text{cluster}_5(x_2) + \text{cluster}_3(x_2)) * P((1,3) (1,3)) = 0.6$ $\text{avg}(\text{cluster}_1(x_3) + \text{cluster}_5(x_3) + \text{cluster}_3(x_3)) * P((3,1) (1,3)) = 0.45$ | X_3 (3,1)  |

Figure 4.7: Simplified (and constructed) pipeline example. The description “here a red T” with gesture at point (1, 3) is represented and mapped to the perceived scenes. Each variant assigns a higher probability to the correct scene, represented by X_2

4.4 Applying the represented knowledge

With all the represented knowledge, the final score for a given candidate scene is computed as follows: for each possible mapping of description object IDs to computer vision object IDs, a gestural score is computed as described in Section 4.2.3 ; the representation of each description is applied to its corresponding object using the grounding just explained; this is combined into an average description score, which is weighted by the gesture score to yield the final score of this mapping for this candidate scene. Figure 4.7 shows a simple example of how each variant processes for a description of a single object, constructed with a simplified coordinate system for the gesture. Each variant is applied to the three candidate scenes on the right side of the figure.

4.5 Experiment

In this section, I describe how I evaluate the overall system performance and the three representation variants.

The evaluation here focuses on the evaluation of the three variants of the representation methods and constrains the uncertainties caused by other modules of the system (e.g., utterance segmentation and gesture recognition). Hence, instead of evaluating the system with the *Multimodal Spatial Description Corpus* described in Chapter 3, we constructed a small spatial scene description corpus, leaving the more complex modelling and evaluation task to Chapter 5.

The evaluation experiments were conducted under 3 setups: **a)** language only descriptions, **b)** gesture only descriptions, and **c)** multimodal descriptions. Next, I will first describe the data collection procedure and the collected data, then provide details on the experiment design and evaluation results.

4.5.1 A scene description corpus

To elicit natural language descriptions, we generated 25 pentomino scenes as illustrated in Figure 4.1. Native German speakers (students from Bielefeld University who were not involved in the project) wrote down verbal descriptions of the scenes. They were asked to start each object description with the keyword **and**, for example, “*here is [OBJECT DESCRIPTION], and [RELATION] is ...*”. Consequently, the utterances can be segmented with a rule-based model and eliminate potential uncertainties of utterance segmentations in the evaluation. As aforementioned, with this data, we simulate the ASR output with the collected verbal descriptions to focus on the core model for this chapter.

In total, we collected 50 scene descriptions. Example (3) shows a sample description of the corpus:

- (3) a. $|_{NS}$ Hier ist $|_{NObj}$ ein pinkes z-ähnliches Zeichen und schräg rechts unten davon ist $|_{NObj}$ ein zweites pinkes z-ähnliches Zeichen und schräg rechts unten davon ist $|_{NObj}$ ein blaues L
- b. $|_{NS}$ here is $|_{NObj}$ a pink Z and diagonally to the bottom right of it is $|_{NObj}$ a second pink Z and diagonally bottom right of it is $|_{NObj}$ a blue L

where $|_{NS}$ indicates the start of a **New Scene** description, while $|_{NObj}$ indicates the start of a **New Object** description.

Similarly, we simulated the gesture detector by taking actual positions of the described objects in the scenes as gesture positions, then adding normally distributed noises to simulate uncertainties of the gesture detector module.

4.5.2 Evaluation

To evaluate the overall performance of the system and the three representation variants, we created a set of test scenes for each scene description. Each test set includes the target scene and 5 other scenes as distractor scenes. The distractor scenes were randomly selected from the set of 25 scenes. This results in 50 test retrieval tasks for each of the 3 variants.

To evaluate the contribution of each modality as well as their joint performance, we designed 3 experiments:

- In Experiment 1, the distractor scenes were modified so that all the objects have the same spatial layout. Therefore, in such cases, gestures, which provide spatial layout information, cannot contribute to the retrieving task.
- In Experiment 2, the spatial layout of objects are kept, but their visual features are modified to be identical. Therefore, language, which describes visual features, cannot contribute to the retrieving task.
- In Experiment 3, both the spatial layout and visual features are kept to evaluate the joint performance of deictic gestures and language descriptions.

In each experiment, we first run the pipeline to build the representations of the descriptions (i.e., 3 sets of representations corresponding the variants A-C), then use the built representations to retrieve the described scene from the set of candidate scenes. Following the evaluation convention of scene retrieval tasks, we evaluated our models with following metrics: **accuracy**, the ratio of correct retrievals; **mean reciprocal rank (MRR)**, which is computed as follows:

$$\text{MRR} = \frac{1}{N} \sum_{i=1}^N \frac{1}{\text{rank}(i)} \quad (4.6)$$

in our setup, the MRR ranges from 1/6 (worst) to 1 (ideal case).

4.5.3 Results

Table 4.2 shows the experiment results of each representation variant under 3 different setups.

In **Experiment 1**, as gestures don't contribute discriminative information, the system performance reflects the contribution of language descriptions in the scene retrieval task. The results show that when only language contributes to the retrieval task, the representation variants can already achieve good performance, with **Variant A** (verbatim representation/word classifiers) having a slight edge on **Variant C** (representation through clustering). Going just with the gesture information as designed, performs on chance level here.

| | | Experiment 1 | | Experiment 2 | | Experiment 3 | |
|----------------|---|--------------|-------------|--------------|-------------|--------------|-------------|
| | | ACC | MRR | ACC | MRR | ACC | MRR |
| Gesture | | 0.1 | 0.37 | 0.65 | 0.75 | 0.67 | 0.78 |
| Gesture+Speech | A | 0.82 | 0.90 | 0.70 | 0.78 | 0.84 | 0.91 |
| | B | 0.68 | 0.81 | 0.68 | 0.76 | 0.68 | 0.81 |
| | C | 0.80 | 0.89 | 0.76 | 0.82 | 0.84 | 0.92 |

Table 4.2: Results of the Experiments. Exp. 1: objects in same spatial configuration in all scenes (per retrieval task); Exp. 2: objects potentially in different configurations in scenes, but same three objects in all scenes; Exp. 3: potentially different objects and different locations in all scenes.

In **Experiment 2**, only gestures contribute to the decision making of the scene retrieval task. All three variants perform robustly only with gesture information: In many sets, gesture information alone already identifies the correct scene (top row, “deictic gesture in the top area of the gesture space”). Language can improve over this in cases where gestures alone compute the wrong mapping of description IDs and object IDs.

In **Experiment 3**, both language and gestures contribute to the scene retrieval task. That is, we evaluated the system performance with the randomly selected test set. **Variants A** and **Variants C** show performances that are much better than **Variant B**. Note that **Variant B** suffers from data sparsity: such as in the training data, the shape U, which is the preferred description in our test data, leading to the wrong shape property being predicted. Interestingly, “compressing” the information into a small set of clusters (in this case, 26) doesn’t seem to hurt the performance.

4.6 Summary

In this chapter, I have presented a system that models the task of understanding multimodal spatial descriptions. The system takes natural language and deictic gestures as input, represents the input information with symbols that are perceptually grounded to real-life scene images. The represented knowledge can be applied to discriminative tasks such as scene retrievals.

This chapter depicts three variants of representing the verbal descriptions: from not compressing the description at all (storing sequences of (stemmed) words, as they occurred), over using pre-specified property symbols to learning a set of “concepts” automatically. In terms of the overall system performance, the system performs well with verbal descriptions. With gesture information providing a large amount of information, the system performance was further

improved.

5

Towards real-time understanding of multimodal spatial descriptions

In the previous chapter, I described a system of understanding multimodal spatial scene descriptions and explored three variants of representing the descriptions. In this chapter, I present a real-time system that understands spatial descriptions incrementally. While the presented real-time system shares the same framework with the one described in the previous chapter, incremental processing models for individual system components are implemented and evaluated in this chapter. Moreover, this chapter introduces the modelling of the spatial description understanding task with the collected spatial description corpus (described in Chapter 3) from humans, hence, the modelling and evaluation are conducted in a more realistic setup. This chapter concludes that abstract deictic gestures not only improve the overall performance of the system, but also result in earlier final correct system decisions.

5.1 Real-time understanding of spatial scene descriptions

As I have mentioned in Chapter 3, psycholinguistic studies show that humans process speech and gestures jointly and incrementally (Campana et al., 2005). While the descriptions unfold, listeners immediately interpret and integrate information from co-occurring speech and ges-

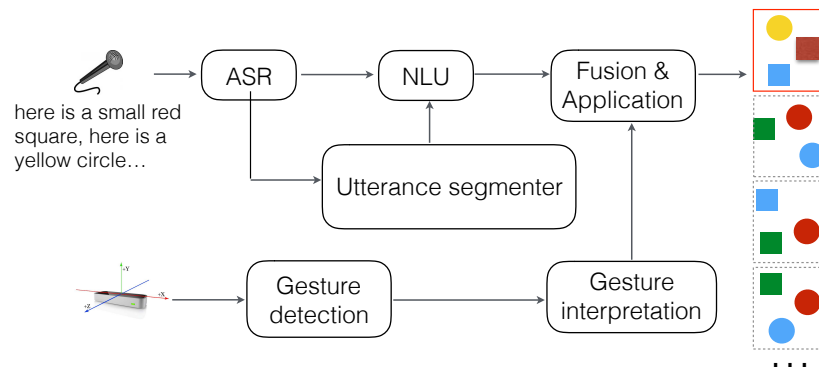


Figure 5.1: Overview of the system.

tures. Moreover, to apply the interpretation later, it's essential to form a hypothesis in mind, making it a very demanding cognitive, language-related tasks (Schneider and Taylor, 1999) for listeners.

To enable a computer system to understand spatial descriptions in the way humans perform the task, it's essential that the system can incrementally process the input information. That is, while a human is talking, the system should be able to understand the description while the description unfolds, and probably make a decision or a clarification request once the important information has been given, rather than reacting only after a full description being given. Moreover, the parallel nature between speech and gestures could also potentially benefit the efficiency of the understanding task. As gestures often precede accompanied verbal content, gestures may benefit the understanding task by providing information earlier than the verbal content in an incremental processing setup.

This chapter illustrates how to model the task of jointly and incrementally interpreting multimodal spatial descriptions, using a simplified spatial description task as described in Chapter 3. Specifically, I will address following questions: **1)** to what degree can deictic gestures improve the overall interpretation accuracy; **2)** how gestures benefit the interpretation of spatial descriptions on the incremental level.

5.2 System overview

As shown in Figure 5.1, the architecture of the real-time system is the same to the one presented in previous chapter (Chapter 4), composed of a speech processing pipeline, a gesture processing pipeline and an application module which applies the interpretations to perform the

scene retrieval task. However, the individual system components are incremental processing models.

Language processing pipeline The language processing pipeline is supposed to interpret natural language descriptions incrementally (i.e., produce understanding results while the description is going on). An automatic speech recogniser (ASR) transcribes speech to words on a word-by-word level. A natural language understanding (NLU) module incrementally processes the output from the ASR module. It sends the latest interpretations to the multimodal application module. Currently, we have only evaluated the system in a simulated setup where manual transcriptions were played back to simulate the ASR to eliminate possible delays and noisy texts caused by ASRs (see Section 5.3 for details). The multimodal application module applies the interpretation results from the NLU module to perform a scene retrieval task also in an incremental manner.

An utterance segmenter takes words from the ASR module, while the words are coming in, it predicts the end of an object description. Instead of manually labelling the end of object description as in previous chapter, we implemented an LSTM model for the segmentation task, which will be described in Subsection 5.2.3.

Gesture processing pipeline In this chapter, I introduce the gesture processing pipeline that is built in a more realistic way: the pipeline takes hand motion data as input and detects the hand positions of deictic gestures, rather than simulating gesture positions with object positions in scenes. In other words, the task of the gesture processing pipeline is to log raw hand motion features from a sensor, recognise gestures and interpret the gestures incrementally (on a frame-by-frame level). While in previous chapter, the simulated deictic gestures were interpreted by comparing its spatial layout with the layout of objects in scenes, the interpretation only happens at the end of a description (i.e., non-incremental). In this chapter, deictic gestures are interpreted incrementally. With each detected deictic gesture, the system first computes which object the gesture is likely to refer to, then compares the spatial layout of detected gestures with corresponding objects in candidate scenes.

After interpreting the gestures, the gesture interpretation module sends the interpretation to the multimodal application module. The pipeline contains 3 components. First of all, a **Leap sensor** tracks hand movements in the effective tracking area. It sends raw hand motion features (i.e., hand palm magnitude, hand palm direction, etc.). A **gesture recognition** (stroke hold recognition) takes the raw features and sends the recognition result (stroke hold position, this case) to a **gesture interpretation** module. The gesture interpretation module processes the gesture information and sends the interpretation to the application module.

Multimodal fusion & application The fusion module takes interpretations from the speech and gesture processing pipelines. Its task is to apply the interpretation, compare the hypothesis with each candidate scene, then select the most likely one as the candidate scene. As the language and gesture processing pipelines work incrementally, the fusion and application modules also work incrementally. Hence, the system can make and refine its decisions while descriptions unfold.

5.2.1 Gesture detection

The task of the gesture detection module is to detect abstract deictic gestures with given hand motion data frames and informs the gesture interpretation module the hand positions of the deictic gestures. We frame the task as a binary classification problem which classifies hand motion frames as stroke hold frame or non-stroke hold frame.

We used labeled hand motion frames to train an LSTM classifier. For each labeled data frame, we selected a time window of 200 ms before current frame. The data frames in the window is composed as a sequence input for each classification task of the classifier. To reduce the input load of the classifier, every other frame in the window was dropped. Following features for each sampled frame were provided to the classifier (in total, each data frame contains 92 features):

- **hand velocity:** the speed and movement direction of the palm in millimetres per second (3 features)
- **hand direction:** the direction from the palm position toward the fingers (3 features)
- **palm normal:** a vector perpendicular to the plane formed by the palm of the hand (3 features)
- **palm position:** the centre position of the palm in millimetres from the Leap Motion Controller origin (3 features)
- **grab strength:** strength of a grab hand pose which ranges from 0 to 1(1 feature). Provided by Leap sensor.
- **finger bone directions:** the direction of finger bones (60 features)
- **finger bone angles:** “side-to-side” openness between connected finger bones (15 features). Provided by Leap sensor.
- **finger angles:** the angles between two neighbouring fingers (4 features).

As aforementioned, stroke holds are featured with low hand velocity. Although it might seem that velocity itself would provide sufficient information for the detection task, we found

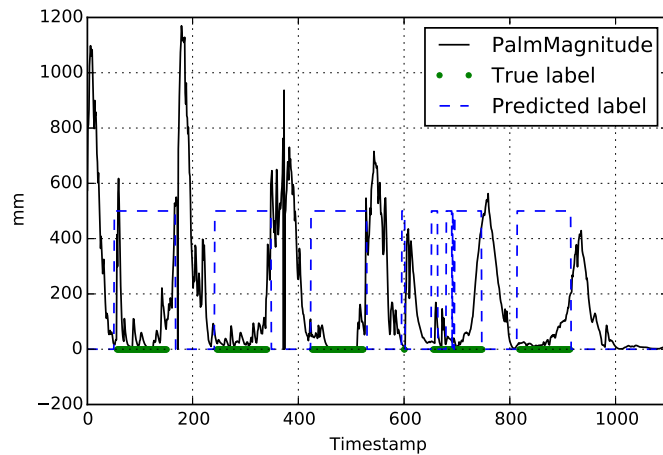


Figure 5.2: Examples of stroke hold detection. We used `palm magnitude` to show the stroke hold phase as it is one of the major features which distinguish stroke hold from other hand movements.

that sometimes participants placed their hands in the gesture space without referring to any object. Using velocity alone for the classification task would cause many false positives. We observed that when hands are not for gesturing, they are usually in a relaxed state with hand palms downward and smaller angles between fingers. In addition, when hands are retrieved to a relaxed state, hands usually switch from gesturing state to relaxed state. In contrast, before gesturing, hands usually switch from relaxed state to gesturing state. Hence, hand frames with a low velocity after a retrieving phase are typically non-stroke holds. Considering these observations, we extracted related the raw features from the hand motion data and trained an LSTM classifier for the classification task.

Architecture of the LSTM The LSTM network includes two hidden layers and a sigmoid dense layer to give predictions. The first hidden layer has 68 nodes and produces 38 tanh nonlinearity. The second layer has 38 nodes and outputs via the dense layer. A dropout layer is applied to the second layer to enable more effective learning. To avoid overfitting, 50% of the input units were dropped and set to 0. The loss function of the model is a binary cross entropy loss function. It was optimised with an rmsprop optimiser. The training was stopped when validation loss stopped decreasing. Figure 5.2 shows some examples of the stroke hold detection.

As a stroke hold often contains several frames, we take the average hand position of all

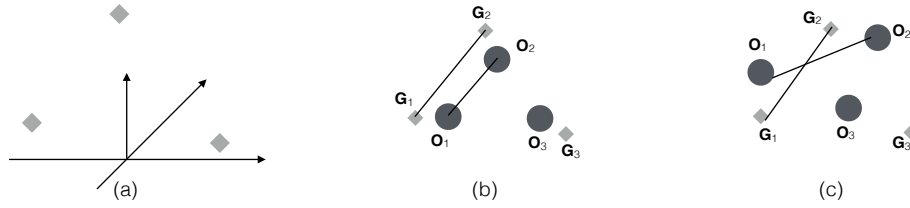


Figure 5.3: Mapping deictics from gesture space to scene coordinate system. (a) deictic gestures in the gesture space (Leap sensor coordinate system); (b) gestures are mapped to the target scene; (c) gestures are mapped to a distractor scene with different spatial configurations.

available frames as the position of the stroke hold. As a result, we can compute and update the stroke hold position while the stroke hold is still going on.

5.2.2 Gesture interpretation

With the position and timestamp of each stroke hold from the gesture detection module, the gesture interpretation module resolves the references of deictic gestures and evaluates how well the spatial configuration of deictic gestures fits with the spatial configuration of a scene image.

Following McNeill’s model (McNeill, 1992), we view the deictic gestures in our task as a reflection of the scene that participants saw on the screen. Namely, they stored the spatial information in their mind as a mental image, then describe the scene with speech and gestures – mapping the mental image to the gesture space. Hence, the spatial configuration of the deictic gestures should reflect the spatial configuration with the one they watched (the target scene that the system should retrieve). For example, a deictic gesture at bottom left of the gesture space is more likely to refer to an object at the bottom of a scene image.

Although the Leap sensor provides hand positions in the 3D space, for the 2D scene description task in this work, we only used the position information in the x and y panel. Therefore, we represented a speaker’s gesture space as $\{(x, y) \in \mathcal{R}^2 : x_{min} \leq x \leq x_{max}, y_{min} \leq y \leq y_{max}\}$, where x_{min} , x_{max} , y_{min} and y_{max} are the boundaries of the gesture space which was estimated from all of a speaker’s gestures.

As aforementioned, the gesture detector sends stroke hold position information from the Leap sensor to the gesture interpretation module. Hence, the position is computed in the Leap sensor coordinate system. To resolve references of objects in the scene image and compare spatial configurations of gestures and objects in the scene, we mapped gesture positions to the

scene image with a simple method. Given a deictic gesture (x, y) , we mapped it to the image coordinate system $\{(x, y) \in \mathcal{R}^2 : 0 \leq x \leq W, 0 \leq y \leq H\}$, and represented the new coordinate as:

$$G = \left(\frac{W * (x - x_{min})}{x_{max} - x_{min}}, \frac{H * (y - y_{min})}{y_{max} - y_{min}} \right) \quad (5.1)$$

where W and H indicate the width and height of a scene image; $x_{min}, x_{max}, y_{min}$ and y_{max} indicate the boundaries of the scene coordinate system. Figure 5.3 shows an example of mapping deictic gestures to a scene image.

In this spatial description task, humans typically describe the objects in a scene sequentially. Hence, we assume that each deictic gesture is meant to refer to only one object. Therefore, the closer a deictic gesture to an object, the likely that the object is the reference of the deictic gesture. With this assumption, we trained a Gaussian kernel density (KDE) estimation model f to turn the distance between a mapped gesture and an object in a scene into a probability:

$$p(O_i|G) = f(\|\mathbf{G} - \mathbf{O}_i\|) \quad (5.2)$$

The probability indicates how likely the object is the correct referent of the gesture. Details of training and evaluation of the model are described in the evaluation section.

While individual gestures provide information for reference resolution, with more than one gestures, we can compare the spatial configuration of the gestures and objects in a scene, which provides information on how well the spatial layout of the deictics fits with all the objects in the whole scene. As shown in Figure 5.3(b) and (c), the better the gestures and objects fit with each other, the smaller the distance between the two vectors. Given two gestures, we first estimate the most likely referential object for each gesture with the KDE model, then measure the fitness of the gestures and objects with the cosine similarity between the gesture and object vectors. Consequently, with $n(n > 2)$ gestures in a scene description, the probability can be computed as following:

$$p(O_1, \dots, O_n | G_1, \dots, G_n) = \prod_{i=2}^n \prod_{j=1}^{i-1} \frac{(\mathbf{G}_i - \mathbf{G}_{i-j}) \cdot (\mathbf{O}_i - \mathbf{O}_{i-j})}{\|\mathbf{G}_i - \mathbf{G}_{i-j}\| \|\mathbf{O}_i - \mathbf{O}_{i-j}\|} \quad (5.3)$$

When there is only one gesture ($n = 1$), no spatial configuration information is conveyed, therefore we set the probability to 1. With each new deictic gesture, the probability can be computed together with all available gestures. In this way, we incrementally applied gesture information to evaluate how well the gestures fit with a scene.

5.2.3 Utterance segmentation

The task of the utterance segmenter is to incrementally identify the start of new object descriptions. Example (1) shows a sample of the description which contains 3 object descriptions (in other words, 3 segments) and corresponding segmentation boundaries (indicated by $|_{SEG}$).

- (1) (a) Hier ist ein blaues Quadrat in rot $|_{SEG}$ hier ist ein gelber kreis $|_{SEG}$ und hier ist ein klein lila kreis
 (b) here is a blue square in red $|_{SEG}$ here is a yellow circle $|_{SEG}$ and here is a small purple circle

For instance, in Example (1), the utterance segmenter is expected to detect the end of segments as soon as it receives the words “hier” (here) and “und” (and), and inform the NLU module.

In our corpus, there are 3 segments in each scene description. Words like “links (left)”, “hier (here)”, and “und” (and) are predictive for segment boundaries as they often occur at the beginning of a segment. However, due to the variability of natural communication, they could also occur in the middle or end of a segment. For example, segments like “a red circle here” and “circle, on the left, red” both occur in the data (see Chapter 3 for details). Therefore, the classifier must learn over a sequence of words to predict segment boundaries.

Architecture of the LSTM We frame the segmentation problem as a binary classification task. An LSTM network was trained for the segmentation task (also using Keras Chollet (2015)). We encoded each word into a one-hot encoding vector (vocabulary size 266) and fed the vector to the LSTM network. The network is composed of one hidden layer and a sigmoid dense layer that gives the prediction. There are 100 nodes in the hidden layer. A dropout layer was applied to it to enable more effective learning. 30% of the output units from the hidden layer were randomly selected and set to 0 to avoid overfitting. The loss function of the model was a binary loss entropy loss function. It was optimised with a *rmsprop* optimiser. The training was stopped when validation loss stopped decreasing.

5.2.4 Natural language understanding

Given transcribed words from the ASR module, the task of the NLU module is to yield a probability distribution over all objects in a scene image. We adopted a natural language understanding model – the **Simple Incremental Update Model** (SIUM), which was originally described in Kennington et al. (2013). SIUM learns a grounded mapping between aspects of language and aspects of visually perceivable (and representable) objects, formalised as follows:

$$p(O|U) = \frac{1}{p(U)} p(O) \sum_{r \in R} p(U|R) p(R|O) \quad (5.4)$$

$p(O|U)$ is the probability of the referential object O behind the speaker's (ongoing) description (a segment as we mentioned in last section) U . This is recovered using the mediating variable R , which is a set of *properties* which map between aspects of each object, for example each object's colour (out of a set of 6 possible colours), shape (square and circle), size (discretised into small, medium, and big), as well as vertical (top, middle and bottom) and horizontal (left, centre and right) placements. We opted for SIUM because it can update the hypotheses incrementally.

The mapping $p(U|R)$ between properties and aspects of U can be learned from data, which we opted for here, by providing the model with words that were observed as being uttered in reference to a specific object and that object's properties. U is represented by ngrams. During application, $p(U|O)$ can produce a distribution over properties which are marginalised over, resulting in a distribution over candidate objects in a scene. This occurs at each word increment, where the distribution from the previous increment is over time. $p(R|O)$ is a simple model of properties belonging to objects, returning 1 if the object O has that particular property, and 0 if it does not.

With a distribution over all of the objects for each segment, we then take these distributions and combine them with gestures, which will now be explained.

5.2.5 Multimodal fusion & application

The task of the multimodal fusion & application module is to fuse speech and gesture interpretations and apply the hypothesis to retrieve the most likely scene from a set of candidate scenes. The whole fusion procedure includes two steps: first, the probability distributions over objects from each modality, speech and gesture, were combined, resulting in a final distribution over objects for each scene, for each segment. Second, combine these various distributions into a final distribution over scenes and retrieve the one with the highest probability. Figure 5.4 shows a constructed example of how speech and gestures were fused.

Reference resolution

For each segment U , we combine the speech and gesture probability distributions to get a fused probability distribution:

$$p(O_i|U, G) = \lambda_1 * p(O_i|U) + (1 - \lambda_1) * p(O_i|G) \quad (5.5)$$

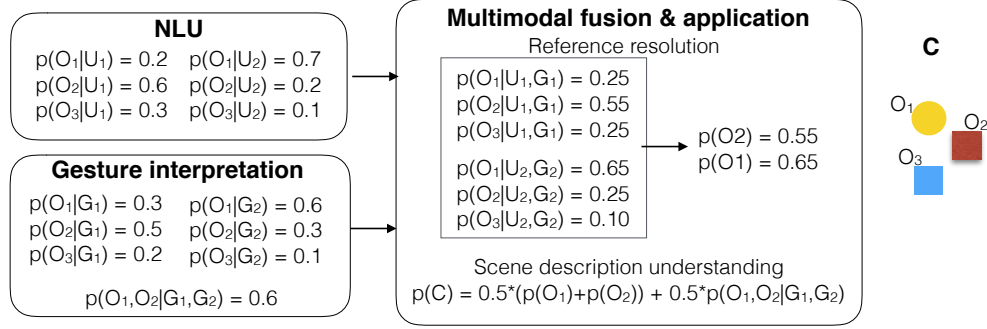


Figure 5.4: Illustration of multimodal fusion & application, given a candidate scene C and following description: U_1 : here G_1 is a small red square, U_2 : here G_2 is a yellow circle. (For clarity of descriptions, the numbers are constructed and not actual computations for this input.)

where λ_1 is a weight parameter. We assume speech and gesture equally contribute to the probability distribution, thus $\lambda_1 = 0.5$ in the setup. When there are no gestures aligned with the segment U , $p(O_i|G)$ was set to 0, hence, there is gestural contribution.

For each segment U , we compute a probability over all objects in a scene. Since in our task, utterances are segmented as descriptions for individual objects, we assume the speaker is referring to one object with a segment. The object with highest probability is taken as the estimated referent for the segment U :

$$O_i^* = \underset{i}{\operatorname{argmax}} p(O_i|U, G) \quad (5.6)$$

The sum of scores for all objects in a scene is taken as the score for the scene.

Scene description understanding

For each candidate scene C , we combined the spatial configuration score with the score from previous steps to get a final score:

$$p(C) = \lambda_2 * \sum_{i=1}^n p(O_i)^* + (1 - \lambda_2) * p(O_1, \dots, O_n|G_1, \dots, G_n) \quad (5.7)$$

the weight parameter λ_2 determines how much the overall spatial layout contributes to the final decision. In our setup, n equals 6. In each retrieving task, the candidate scenes include a target scene and 5 distractor scenes (see 5.3.4 for details).

5.3 System evaluation

We evaluated our system based on the *Multimodal Spatial Description Corpus* described in Chapter 3. The evaluation experiments were conducted with a “hold-one-out” setup. In each evaluation, data from one participant was left as test data while other data as training data to prevent the system from learning about possible idiosyncrasies of a speaker on whom it is tested. In the rest of this section, I will first describe the evaluations of individual system components, then describe whole system performance evaluations and discuss the results.

5.3.1 Gesture detector evaluation

First of all, we evaluated the gesture detector. Following the convention of gesture classification evaluations, we used **F1-score**, **precision** and **recall** as the evaluation metrics.

The gesture detector achieves an **F1-score** of 0.85, **precision** 0.77, **recall** 0.94. Each gesture classification task takes around 10 to 20 ms, correlating to the computational ability of the machine. The reported results here are computed on a MacBook Pro 2015 with following hardwares: processor 2,9 GHz Intel Core i5, memory 8 GB 1867 MHz DDR3.

Currently, we haven’t evaluated other traditional gesture classifiers for the system, thus we haven’t compared the LSTM model with other models. Since the focus of the current evaluation is interpretation and application of the multimodal descriptions, we leave it as future work to implement other models and compare the performance with the current model.

5.3.2 Gesture interpretation evaluation

We evaluated the kernel density estimation (KDE) model of the gesture interpretation module by object **reference accuracy**. Namely, given a deictic gesture position, how often does the referential object gets the highest score among all candidate objects in a scene?

As aforementioned, we mapped deictic gesture and a candidate scene to the same coordinate system, then computed the distances between the mapped gesture and candidate objects in the scene. With the computed distances, we fit a Gaussian KDE model (with the bandwidth setting to 5) using the distances in the training data.

We tested the trained KDE model with computed distances in the test data. The model achieves an average **accuracy** of 0.81, which significantly out forms the chance level accuracy 1/3.

| Model | F1-score | Accuracy | Recall | STD of accuracy |
|----------|----------|----------|--------|-----------------|
| Baseline | 0.84 | 0.85 | 0.83 | 0.23 |
| LSTM | 0.89 | 0.92 | 0.88 | 0.1 |

Table 5.1: Evaluation results of utterance segmenter.

| Metrics | Speech only | Gesture only | Speech + gesture |
|------------------------|-------------------|------------------|------------------|
| Our system | 0.75/ 0.79 | 0.50/0.50 | 0.84/0.85 |
| <i>Human eval</i> | 0.86/- | 0.32/- | 0.77/- |
| <i>Random baseline</i> | 0.17/0.41 | 0.17/0.41 | 0.17/0.14 |

Table 5.2: Results of whole system evaluation.

5.3.3 Utterance segmentation evaluation

We evaluated the utterance segmenter performance with **F-score**, **accuracy** and **recall**. As shown in Table 5.1, the LSTM model outperforms the keyword spotting baseline model. It achieves a higher precision score and a lower standard deviation (SD) of precision. The lower SD indicates stable predictions between participants which is important for real-time systems that are expected to work with users who didn't contribute to the training data. Although currently, the LSTM model only marginally outperforms the baseline model, it's likely that with more training data and more various descriptions, the LSTM model will perform even better.

The NLU component was evaluated as part of the whole system, which will be described now.

5.3.4 Whole system evaluation

To assess the overall performance of our system, we designed "hold-one-out" offline tests with the collected data. We simulated real-time multimodal spatial descriptions by playing back the collected multimodal descriptions in real time. The transcriptions of speech were played back, simulating output of an incremental ASR. Stroke hold positions detected from the motion data were played back as gesture information (hence, the gesture detection wasn't performed in real time, so that currently we don't have to consider the delays caused by the gesture detection module). The tests include mono-modal and multimodal setups: speech only, gesture only and speech+gesture.

The system performance was evaluated with a scene retrieval task. That is, given a multimodal description, the system is supposed to retrieve a scene from a set of candidate scenes

that fits best with the description. We created a test set for each scene in the corpus. Each test set includes the target scene and 5 randomly selected distractor scenes. Hence, the chance level accuracy of the scene retrieving task is 1/6.

The metrics

We evaluated the system performance with **mean reciprocal rank (MRR)** which is computed as follows:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \quad (5.8)$$

where $|Q|$ indicates the set of scene retrieval queries. For each scene retrieval query, we rank the candidate scenes in ascending order according to the scores from the fusion module (e.g., the scene with highest score got a rank of 1). As there are 6 candidate scenes in each retrieval query, the rank ranges from 1/6 (the worst case) to 1 (the ideal case). **MRR** ranges from 0.41 (the worst case) to 1 (the ideal case).

Speech only

Test setup: In this test, only speech contributes information. We replayed audio transcriptions in a real-time pace to simulate the speech only descriptions. The weight parameters λ_1 and λ_2 in Equation 5.5 and Equation 5.7 were set to 1.0.

As shown in Table 5.2, the average **MRR** of the tests is 0.79 (accuracy 0.75), which significantly outperforms the baseline. Although the evaluation setups are the same for all participants, when comparing evaluation score of each participants, we observed individual differences between participants. The difference could be due to varied language descriptions, such as referring to the same colour or shape with different words or omitting spatial descriptions in verbal descriptions (see Chapter 3). The varied language descriptions affect the utterance segmenter performance as well as the NLU module performance. Hence, they consequently affect the general system performance.

Gesture only

Test setup: In this test, we only replayed the hand motion data to simulate the gestures. The weight of gesture information in fusion module (Equation 5.5 and Equation 5.7) were set to 0, so that only gestures contribute to making the scene retrieving decision.

The average **MRR** of all tests is 0.50 (accuracy 0.50). It outperforms the chance level baseline MRR by 0.09, which underperforms the multimodal model and the language only model. On one hand, gestures are ambiguous, as the spatial layout encoded in gestures is

approximately mapped to the space from human mind. On the other hand, gestures only convey positional information of referents and relative spatial configurations of referents, the similarity between targeted scene and distractors also affect the results.

Speech + gesture

Test setup: In this test, we replayed speech transcriptions and hand motion data in parallel to simulate real-time multimodal descriptions. λ_1 in Equation 5.5 and Equation 5.7 were set to 0.5, hence, speech and gestures contribute equally.

Although as shown in Table 5.2, gestures are less reliable than language descriptions, considering that abstract deictic gestures can also contribute equally by complementing the language when spatial descriptions are omitted, we assumed that speech and gestures contribute equally in all descriptions. It's possible that learning optimised parameters which can adjust to language and gesture descriptions will lead to better performance. We leave it as future work to include such a component in the system.

The multimodal model achieves the best performance among all test, with an average **MRR** of 0.84. It shows that gestures help to improve the performance of the system. Next, I discuss the evaluation results.

Discussion

The system evaluation results described above show that our gesture interpretation method can successfully extract spatial information from gestures. While the language description itself can already achieve a good performance, adding gesture information further improves the system performance, although only marginally.

Hereby, I discuss the reasons of the limited improvement. One possible reason for the limited improvement is that object position information is often redundantly encoded in verbal descriptions. The overlap between speech and gestures has been observed, described and discussed in previous works (Epps et al., 2004). In our case, the data collection may further encouraged such overlaps, as participants were instructed to describe object positions (see Chapter 3). In situated communications, it's less likely that humans would mention all attributes of referents, in which case, gestures would contribute more prominently.

In terms of real-time systems, the system performance is always affected by all individual components. Note that current evaluations are only offline tests, using speech transcriptions to simulate language descriptions. In a realistic setup, the language description should come from an automatic speech recogniser (ASR), which provides noisier words than manual transcriptions. In this case, the redundancy of gestures will disambiguate the uncertainty resulted from

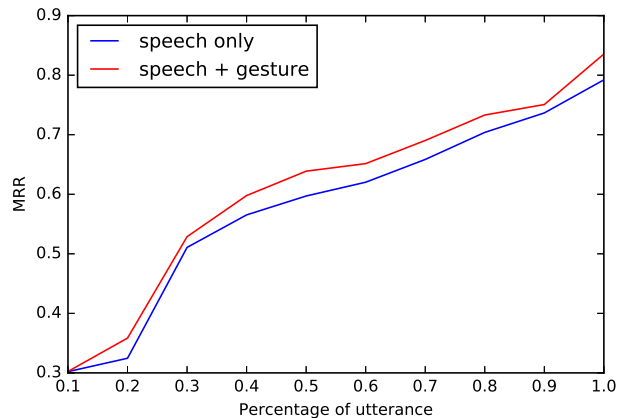


Figure 5.5: Average MRR of incremental evaluation.

the ASR. In the future, I will evaluate our system with live spatial descriptions.

5.3.5 Incremental evaluation

A system being able to incrementally process input information would lead to earlier interpretation than systems that only process information after descriptions end, which is important for real-time systems. Therefore, we evaluated the system performance on the incremental level in speech only and speech plus gesture setups, using incremental evaluation metrics (Schlangen and Skantze, 2009) as follows:

- **average first correct (fc):** how deep into the utterance (as percentage of the whole utterance duration) does the system makes a correct decision the first time, potentially changing its mind again later?
- **average first final (ff):** how deep into the utterance does the system makes a correct final guess?
- **average edit overhead (eo):** ratio of necessary edits/all edits, indicating how stable the decisions of the system are.

As shown in Figure 5.5, incorporating gestures doesn't lead to an earlier first correct decision on average. In both language-only and multimodal tests, the average **fc** is 0.24. It might be due to the fact that descriptions often start with speech, hence speech contributes earlier than gestures. Moreover, at the beginning of a scene description, the first deictic gesture simply

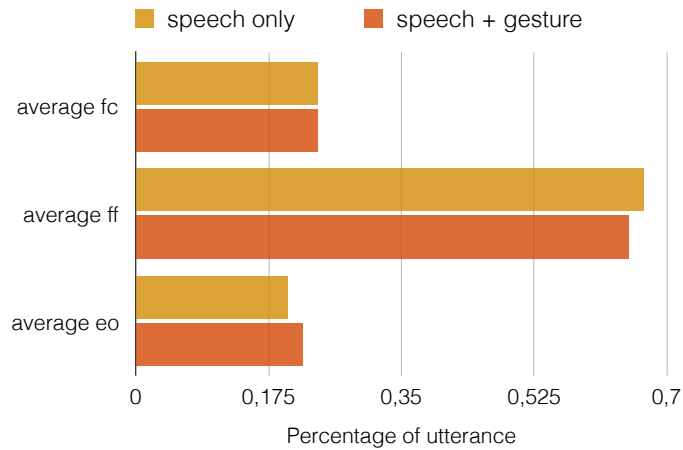


Figure 5.6: Results of incremental evaluation. See text for description of metrics. For all metrics, lower numbers denote better performance.

indicates an object in the gesture space, without positional information of other objects, the gesture is not able to differentiate candidate scenes.

When combining speech with gestures, the average **eo** is slightly higher. It shows that gestures do contribute information in making the retrieving decision. With the complementary information of gestures, the system risks more edits to move toward a right decision. The reward of more edits is an earlier first final correct decision (**ff**).

As shown in Figure 5.5, when combining gestures with language descriptions, the system achieves an average **ff** of 0.65, comparing with a value of 0.67 in speech only situation. For example, when a description lasts for 30 s, incorporating gestures helps the system to achieve a first final decision 600ms, which is noticeable, albeit not in a large scale.

Figure 5.5 plots **MRR** over the course of the utterances (to be able to average, again expressed as percentage of full utterance). **MRR** increases continuously, indicating that for this task, important information can still come late.

Discussion

As shown above, gestures enable the system to achieve earlier first final correct decision, which will benefit situated communications. It's promising that in situated dialogues, the system will understand spatial descriptions without waiting for completed verbal descriptions, making a computer system more human-like. Moreover, the overhead edits caused by gestures can also contribute to more interactive system feedbacks. For example, when the system “realises”

that the system's decision becomes more ambiguous (in other words, leading to bad retrieving decisions), the system can request for clarifications so that the route giver can timely change the description strategy to make the understanding task easier to the system. Such interactions will lead to more efficient human-system interactions.

5.3.6 Human understanding

To ground our results in human performance, we also evaluated human performances of understanding the spatial scene descriptions. We randomly selected 65 scene descriptions (5 scene descriptions from each participant) and asked workers from the Crowdfunder platform to select the described scenes from 5 distractor scenes, the same as the evaluation setup of the real-time system.

Similar to the real-time system evaluation setup, we asked workers to select described scenes with multimodal descriptions, language-only descriptions and gesture-only descriptions. In the language-only setup, the workers beat our system with an accuracy of 0.86. With the gesture-only descriptions, our system outperforms human workers. The accuracy is 0.18 higher than the human performance. Interestingly, when given multimodal description, the gestures seem to be distracting to the workers, who performed not only worse than our system, but also performed worse than when only giving language descriptions. This is presumably due to the heavy cognitive load of observing the gestures and evaluating the scenes.

The less well human performance with multimodal descriptions suggest that, in real interactions, the delivery of such descriptions would be much more interactive and delivered in instalments. An instruction giver might adjust her/his descriptions according to the listeners' interactive feedbacks or clarification requests. To model this interactive instruction giving and understanding procedure, a system not only needs to interpret the descriptions incrementally, but also needs to interpret listeners' feedbacks. We leave it as future work to model interactive spatial descriptions.

Discussion

On the general performance level, the evaluation results show that our method can successfully extract spatial information from the gestures. The "speech-only" condition also achieves good performance. When combining speech and gestures, the system performance was further improved, which is consistent with our expectation. However, the improvement is somewhat limited. One reason for this is that position information is often redundantly encoded in verbal descriptions. Overlap in content between gestures and speech has been observed in previous work (Epps et al., 2004); the experimental setting may have further encouraged such redundant

encoding. In real situations, it may be less likely that speakers indeed mention all attributes, in which case contributions of modalities may be more complementary. (The system, in any case, would be ready to handle this.) In a practical system, this redundancy might even be a useful feature. Here, we allowed the system incremental access to the manual transcription of the speech. In a live system, this information would come from automatic speech recognition (ASR), and would be more noisy. We speculate that the redundancy coming from the gestures will then help locally disambiguate the ASR output. We will test this in future work.

On the incremental level, gestures help to achieve an earlier correct final decision. It's promising that in situated dialogues, the system might understand descriptions from humans without waiting for all verbal descriptions and thus may behave more human-like. Moreover, gestures result in more overhead edits (Figure 5.5). This signal can be used for clarifications in situated dialogues. For instance, while a route giver notices that the system's decision changes to bad decisions, the route giver might change the description strategy to make the decoding task easier, or the system can make clarification requests. We believe these signals will lead to more human-like interactions.

5.4 Summary

In this chapter, I described a multimodal system that builds and applies spatial descriptions to a scene retrieval task. The evaluation results in uni-modal setups show that both speech and gestures are informative for the scene retrieval task. Combining speech and gestures further improves the system performance. Furthermore, the system was evaluated in terms of incremental processing performance. The results show that gestures help to achieve earlier final correct decisions. Hence, gestures not only contribute information, but also benefit interpretations on the incremental level due to its parallel nature with speech. This will benefit more dialogical tasks such as triggering immediate clarification request. I leave it as future work to implement more dialogical systems.

6

Investigate symbolic and iconic modes in object descriptions

The previous two chapters described a real-time system that models the representation and real-time incremental processing of multimodal spatial descriptions composed of abstract deictics and verbal utterances. In this chapter, I investigate the interplay of semantics between symbolic (natural language) and iconic (hand-drawn sketches) modes in multimodal object descriptions. Based on the *Multimodal Object Description Corpus* described in Chapter 3, the meanings of multimodal object descriptions are modelled; the contributions of natural language and sketches with an image retrieval task are evaluated. I show that multimodal descriptions outperform verbal- or sketch-only descriptions. Adding even very reduced iconic information to a verbal image description improves the image retrieval accuracy.

6.1 Draw and Tell: iconic and symbolic modes in object descriptions

In natural interactions, descriptions are typically multimodal: we describe a route as “along the fountain” while gesturing the trajectory of the route into the air or sketch the trajectory on a piece of paper (Emmorey et al., 2000a; Tversky et al., 2009); we may describe an elephant as

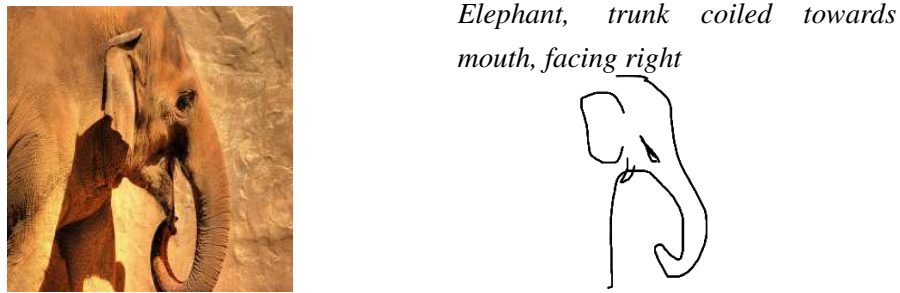


Figure 6.1: A photograph; a verbal description of its content; and a sketch.

“facing right, trunk coiled towards to mouth” and gesture/sketch how exactly the trunk is coiled, as shown in Figure 6.1. While natural language conveys symbolic information conveniently (e.g., describing the *category* of a landmark or an object with words *fountain* and *elephant* or describe *visual attributes* as “facing right”), iconic information encoded in gestures/sketches convey information visually, indicating visual features that are difficult to encode in language, such as the exact the shape of a *coiled trunk*.

Especially, descriptions of visual objects or situations can be supported by the iconic mode of reference provided by gestures or sketches, that is, reference via similarity rather than symbolic convention (Pierce, 1867; Kendon, 1980b; McNeill, 1992; Beattie and Shovelton, 1999). However, in previous work, these descriptions are typically ‘mono-modal’, either using purely verbal descriptions (Schuster et al., 2015; Hu et al., 2016),¹ or via hand-drawn sketches (Sangkloy et al., 2016; Qian et al., 2016; Yu et al., 2016).

I’m interested in modelling the joint contribution of symbolic (i.e., natural language) and iconic (i.e., hand-drawn sketches) modes. A direct, but controlled model of this task is the image retrieval task that retrieves one photograph out of many photographs based on a description of it. With the Multimodal Object Description corpus introduced in Chapter 3, I investigate the interplay of semantics of symbolic and iconic mode in multimodal object descriptions. In particular, the rest of this chapter addresses following research questions: **a)** How well natural language and hand-drawn sketches perform on their own in an image retrieval task (i.e., mono-modal models); **b)** How well the iconic information in sketches can improve the performance of natural language descriptions in an image retrieval task (i.e., multimodal model); **c)** In order to be informative, how many details should be included in a sketch?

In what follows, I first describe how the meanings of multimodal object descriptions are modelled, then describe and discuss the evaluation experiments.

¹As implemented and in commercial use on popular internet search engines.

6.2 Model the meaning of multimodal object descriptions

Using a model of grounded language semantics and a model of sketch-to-image mapping, we model the multimodal meaning of object descriptions. We compose a joint meaning representation out of individual description components, namely, verbal descriptions and hand-drawn sketches. Then we combine the meanings from both modalities using a late fusion approach (Atrey et al., 2010).

6.2.1 Grounding verbal descriptions

To provide an objective judgement on how well a verbal description fits with a photograph, we adopt the “words-as-classifiers” (WAC) model (Kennington and Schlangen, 2015; Schlangen et al., 2016), grounding a verbal description to visual features of a photo.

The WAC model trains a logistic regression classifier for each word. The classifier takes the feature vector of a photo as input, and provides a score which reflects the fitness between the photo and the word (0 the worst and 1 fits perfectly). The feature vectors of photos were extracted with a convolutional neural network (i.e., the Triplet Network model in this work) trained on the Sketchy dataset (Schroff et al., 2015; Sangkloy et al., 2016).

Each word classifier was trained with feature vectors of all images that were described with the word. For example, to train a classifier for the word *elephant*, we take all the images that described with the word *elephant*, using them as positive features, while randomly select the same amount of images that were not described with *elephant*, using them as negative features. As some words are rare in the corpus (e.g., “smiley”), we only trained classifiers for words which appear more than 10 times.

Given an image description $D : w_c, w_{a_1}, w_{a_2}, \dots, w_{a_i}$, where w_c indicates the category word of an object and w_{a_i} indicates an attribute word in the object description, we compute a combined fitness score between the description D and a candidate photograph \mathbf{P} as following:

$$s_D(D|\mathbf{P}) = s_{w_c}(\mathbf{P}) \times \sum_{i=1}^n s_{w_{a_i}}(\mathbf{P}) \quad (6.1)$$

where $s_w(\cdot)$ indicates the classifier score. Attribute scores are additive, while the overall attribute score and the category score are multiplicative.

For some attribute words that are in the test set but not in the training set, we simply left out the words in the evaluation process.

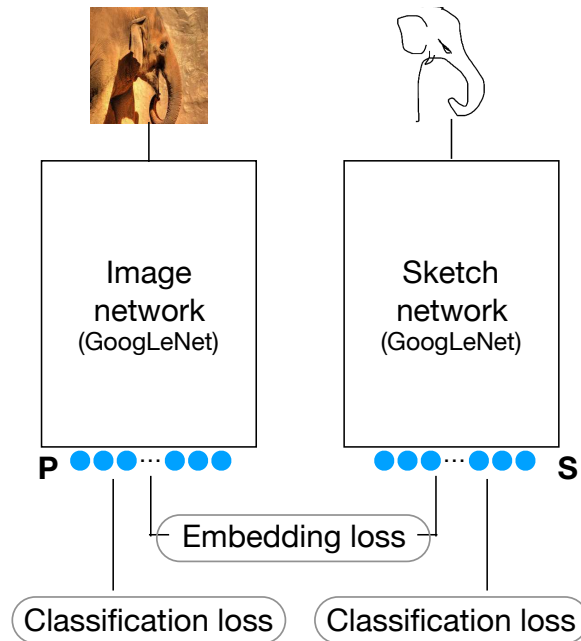


Figure 6.2: The GoogLe network from Sangkloy et al. (2016). The *Image network* and the *sketch network* are both pre-trained with an image/sketch classification task (with classification losses), then fine-tuned for an sketch based image retrieval task with an embedding loss. **P** and **S** indicate the feature vectors that represent images and sketches. For detailed descriptions of the network and the training procedure, please refer to the original paper.

6.2.2 Comparing sketches with images

The original work that introduced the Sketchy dataset provided a “Triplet Network” model that embeds images and features into the same vector space (Sangkloy et al., 2016). As shown in Figure 6.2, the network is composed of a *sketch net* and an *image net*, encoding sketches and photos into vectors in a shared vector space.

Hereby, I briefly describe the “triplet network”. For more details, please refer to the original paper. The “triplet network” is composed of a sketch network and an image network, which are with the same structure. The whole network was trained with a ranking loss function, with input tuples of the form (S, I+, I-) corresponding to a sketch, an image that matches the sketch and an image that is non-matching. By minimising the loss function, sketch and corresponding photos are mapped into the embedding space as close as possible to each other. We take the output feature vectors of the final layer in sketch and image networks, and use the feature

vectors to represent sketch and images.

As all the images and sketches are mapped to a joint embedding vector space, the vector distances between sketches and images reflect their similarities. A smaller distance indicates better visual similarity. Hence, we turn the reciprocal of the distance to a score to measure the fitness between a sketch and an image, as follows:

$$s_{sk}(\mathbf{P}|\mathbf{S}) = d(\mathbf{P}, \mathbf{S})^{-1} \quad (6.2)$$

where \mathbf{S} and \mathbf{P} correspondingly indicate feature vectors of a sketch and a photograph derived with the sketch network and the image network.

6.2.3 Fusion

We multiplied the scores of sketches and verbal descriptions to derive a final score, making it a late fusion approach as follows:

$$s_{sk+cat+att} = s_{sk}(\mathbf{P}|\mathbf{S}) \times s_D(D|\mathbf{P}) \quad (6.3)$$

6.3 Experiments

In this section, I provide the details on how we evaluated the proposed models. To inspect the contribution of language and sketches in image descriptions, we evaluated the performance of the model with following setups: verbal descriptions, sketch-only and verbal descriptions with various ratios of sketch details, as shown in Table 6.1. We will first describe the evaluation metrics, then give details on and discuss evaluation results of mono-modal models and multimodal models, as well as evaluations results with reduced sketches.

6.3.1 The image retrieving task

We evaluated the interplay of symbolic and iconic semantics with an image retrieving task with mono-modal and multimodal setups. Given a verbal description and a sketch of a photograph, we aim to retrieve the target image from a set of the images.

We conducted the experiments on the Sketchy data Sangkloy et al. (2016). The original data was split into train-test sets for training and evaluating the sketch-image network. To use the sketch network, we followed the original train-test setup. In total, there are 1071 images in the test set. Each image in the test set was paired with around 5 sketches (i.e., 5375 sketches in total). The chance level accuracy for the image retrieving task is 0.000933.

6.3.2 Metrics

Following the conventions of evaluating image retrieving tasks, we measured the performance of the image retrieving models by average Recall@K.

For each retrieving task, if the target image is ranked among the top **K** candidate images, the recall @K equals 1, otherwise, recall @K equals 0. We take the average recall over all retrieving tasks to measure the performance of a model. In Table 6.1, we report the average recall of @K=1 and @K=10 for each evaluation setup.

6.3.3 Experiment 1: Mono-modal models

First of all, we evaluated the mono-modal models to quantify the performance of each modality. Namely, verbal only model and sketch only model (*sk*). For the verbal only evaluation, we further evaluated the category word only model (*cat*), the attribute words only model (*att*), intended to inspect the contribution of words in terms of distinguishing objects within and across categories.

Category word only (*cat*) Category words make objects discriminative among objects from other categories.

For category word only evaluation, we judged the fitness between an image and a category word with the WAC model. Among all the candidate images in the test set, the one with highest score is retrieved. The model achieves an average recall of 0.12 (@K=1) and 0.9 (@K=10), which is higher than a random chance level recall @K=1 of 0.093%, but much lower than the ideal score 1.0.

As in the test set, there are 86 images in each category, hence, given the category word, the chance level recall is 0.12 (@K=1). It shows that the model can efficiently distinguish objects between categories, but not within categories.

Attribute words only (*att*) Attribute words describe the colour, shape, orientation and other visual attributes. Attribute words can be discriminative among objects which are within the same category and with different visual attributes (e.g., an elephant “facing right” vs. an elephant “facing left”). However, objects from different categories might share the same attributes, thus attribute words are not discriminative for objects from different categories. For example, an elephant and a bear can both be described with the attribute “facing right”. Hence, the visual attribute “facing right” is not discriminative.

In this setup, we take all the attributes of each object description, compute a fitness score using the WAC model for each attribute word, then take the average fitness score as a final

fitness judgement. The model achieves an average recall of 0.03 (@K=1) and 0.23 (@K=10).

The results are not surprising. Given that the attribute words such as “facing right” can be used to describe many images both within and across categories. The attribute words are not very discriminative on their own.

Category word + attribute words (*cat+att*) Category words and attribute words jointly delineate an object within and among object categories. Therefore, we combined category words and attribute words for a full verbal evaluation model.

With this setup, we simply take the average score of category words and attributes for the fitness judgement. The model achieves an average recall of 0.14 (@K=1) and 0.83 (@K=10). While the joint model outperforms the *cat*- and *att*-only models in terms of Recall @1 score, it underperforms the *cat*-only model in terms of Recall @10 score. We conjecture, this is due to the fact that objects among different categories can be described with similar or the same attribute words. Therefore, attribute words may bring in noisy information when distinguishing object across categories, where category words *cat* is informative.

Sketch only (*sk*) We also evaluate the sketch only model with various ratios of sketch details, ranging from 10% to 100%.

As shown in Table 6.1, the average recall increases with the increase of the ratio of sketch details. Looking more closely, at the beginning, the average increases more. This shows that humans often draw most salient parts first, and enrich the sketch with small details later. Given 100% sketch details, the model achieves a Recall @ 1 score of 0.35. While outperforming the language-only models, it underperforms the multimodal models. The results demonstrate that while the iconic information encoded in sketches are informative to a certain degree, a joint retrieving model can benefit from incorporating symbolic information in natural language.

6.3.4 Experiment 2: multimodal models

We evaluated the retrieval performance of the multimodal model, where verbal descriptions and sketches contribute together. Similar to the mono-modal evaluations, we evaluated multimodal models in following setups: *cat+sketch*, *att+sketch* and *cat+att+sketch*.

As shown in Figure 6.1, while full sketches achieve 0.35 of Recall @1 (Recall@10 = 0.84), when combining full sketches with attribute words, the average Recall @1 increases to 0.37 (Recall@10 = 0.87). In comparison, combining category words with full sketches achieves an average Recall@1 of 0.41 (Recall@10 = 0.96). This indicates that, category words are complementary to sketches (iconic information), while attribute words supplement with sketches to a larger extent.

| Sketch Detail Recall | 10% | | 30% | | 50% | | 70% | | 90% | | 100% | |
|-------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | @1 | @10 | @1 | @10 | @1 | @10 | @1 | @10 | @1 | @10 | @1 | @10 |
| <i>sk</i> | 0.01 | 0.06 | 0.07 | 0.27 | 0.17 | 0.55 | 0.25 | 0.70 | 0.31 | 0.79 | 0.35 | 0.84 |
| <i>att</i> | 0.03 | 0.23 | 0.03 | 0.23 | 0.03 | 0.23 | 0.03 | 0.23 | 0.03 | 0.23 | 0.03 | 0.23 |
| <i>cat</i> | 0.12 | 0.90 | 0.12 | 0.90 | 0.12 | 0.90 | 0.12 | 0.90 | 0.12 | 0.90 | 0.12 | 0.90 |
| <i>cat+att</i> | 0.14 | 0.83 | 0.14 | 0.83 | 0.14 | 0.83 | 0.14 | 0.83 | 0.14 | 0.83 | 0.14 | 0.83 |
| <i>sk+att</i> | 0.03 | 0.16 | 0.09 | 0.39 | 0.20 | 0.64 | 0.28 | 0.76 | 0.33 | 0.83 | 0.37 | 0.87 |
| <i>sk+cat</i> | 0.12 | 0.76 | 0.20 | 0.85 | 0.28 | 0.92 | 0.34 | 0.94 | 0.38 | 0.96 | 0.41 | 0.96 |
| <i>sk+cat+att</i> | 0.15 | 0.81 | 0.21 | 0.87 | 0.30 | 0.92 | 0.35 | 0.94 | 0.38 | 0.95 | 0.41 | 0.96 |

Table 6.1: Average recall at K=1 and 10, at different levels of sketch detail. Highest number in column in bold. Numbers for language-only conditions do not change with level of sketch detail.

When combining full verbal descriptions with full sketches, the model achieves a Recall@1 score of 0.41 (Recall@10 = 0.96), which equals the performance of *cat+sk*.

6.3.5 Experiment 3: reduced sketch details

While the sketch dataset was collected in an offline setup, in which participants were allowed to draw full sketches of objects without verbal descriptions in parallel, in situated communications, humans typically speak and sketch simultaneously. Thus, due to timing pressure, humans often only sketch part of an object, rather than give full sketch details. For example, a speaker may only sketch the trunk of an elephant to make it distinguishable from other elephants, while describing colour, orientation and other visual attributes verbally. Therefore, we investigated the performance of the retrieving models with reduced sketch details at 10%, 30%, 50%, 70%, 90% respectively. By systematically reducing sketches and combining the reduced sketches with images, we show to what degree verbal descriptions can recover the information contributed by sketches.

Reduced sketches also simulate situated communications, where humans only draw the most salient part of an object while complementing the sketch with verbal descriptions. We leave it as future work to conduct experiments with salient strokes.

As aforementioned in Chapter 3, the sketches were stored as SVG files, an XML-based vector image format that provides high resolution start and end timing information of each stroke. This allowed us to systematically reduce the sketch details according to their timestamps. For each sketch, we take the first $n\%$ strokes according to their end time. On one hand, this re-

flects the sequential aspects of strokes which sheds light on incremental processing of image retrieving, on the other hand, humans tend to draw salient features first. Hence, we assume that earlier strokes reflect salient features among all strokes of sketches.

As shown in Figure 6.1, the average Recall@1 of the full model (*cat+att+sk*) increases with increased sketch details. 70% sketch with verbal descriptions can achieve the same performance as full sketches. In other words, 30% details of the sketch can be recovered by verbal descriptions.






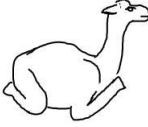



| <i>cat+att</i> | 30% <i>sk+cat+att</i> | 30% <i>sk</i> | 100% <i>sk</i> |
|--|---|---|---|
| chicken, can see head only, head is mainly red skin |  |  |  |
| Rank=1 | Rank=1 | Rank=27 | Rank=1 |
| camel, light brown, laying down, head on right, has blanket to ride on |  |  |  |
| Rank=3 | Rank=1 | Rank=29 | Rank=1 |
| butterfly, facing left, white |  |  |  |
| Rank=3 | Rank=1 | Rank=32 | Rank=1 |

Figure 6.3: Retrieval with verbal description only (1st column), verbal description plus 30% sketch (2nd column), 30% sketch (3rd column) and 100% sketch (4th column).

Figure 6.3 shows three examples of the ranking results. For a chicken image, with 100% sketch or only the verbal description, the target image ranks at top 1 among 1071 candidate images. When reducing the sketch details to 30%, the rank decreases to 27. Combining the

verbal description with 30% sketch ranks the target image at top 1. Similarly, for a camera and a butterfly image, the verbal descriptions rank the target image at 3, when enhancing the verbal description with 30% sketches, the target images are ranked at top 1. Therefore, only a limited amount of sketch details help to improve the retrieval performance.

6.4 Discussion

While we used hand-drawn sketches in this work, the results are also interesting for interpreting iconic gestures. Sketches are similar to gestures in the sense that both signify iconicity in a visual way. As we have seen, sketches can be encoded into feature vectors which represent corresponding iconic information. We believe it's also possible to encode iconic gestures into feature vectors to represent iconic information. The challenge lies in collecting large scale corpora of iconic gestures.

However, gestures are different from sketches in terms of following aspects: 1) Gestures visualise iconicity in a shared physical space where the trajectory disappears immediately. Therefore, the interpretation of gestures must proceed in a time-constrained, incremental manner, so as to not overload a listener's visual memory. 2) Iconic gestures are usually accompanied and synchronised to speech. Hence, gestures cannot provide as many details as sketches do, but only some most salient iconic features of mentioned objects. In other words, they are closer to our reduced sketches. 3) As gestures encode fewer details, the interpretation of gestures can be largely dependent on accompanied language. In comparison, sketches can encode as many details as one intends to, thus the interpretation of sketches are less dependent on verbal content; 4) The person producing the sketch can go back and correct themselves. In contrast, we cannot look at our gestures and re-gesture. Therefore, iconicity in gestures is more abstract, distorted than in sketches. The above challenges make the interpretation of gesture related multimodal communications more challenging than interpreting sketches. We leave it as future work to investigate how to model the meaning of iconic gestures.

6.5 Summary

In this chapter, I introduce a study that investigates the interplay of semantics between symbolic and iconic modes in object descriptions. Combining a model of grounded word meaning with an existing model of image/sketch embedding, I show that multimodal object descriptions outperform verbal- or sketch-only descriptions in an image retrieval task. Furthermore, adding even reduced sketches improves the overall performance of the image retrieving task. The verbal descriptions can make up to 30% of reduced sketches. This suggests that, interfaces

allowing iconic gestural input, which also provide reduced iconic information, may also enable machines to understand humans better than language-alone interfaces.

7

Learning semantic categories of multimodal descriptions

After introducing works on modelling multimodal descriptions that contain natural language and abstract deictics/sketches, in this chapter, I present a study conducted with a more realistic dataset - the SAGA corpus, which contains multimodal route descriptions. The results of the study show that natural language is informative for the interpretation of co-verbal iconic gestures, which convey meanings by assembling visual similarities to referents. Although various works in gesture studies have shown that the interpretation of iconic gestures depends on the accompanied content, for human-computer interfaces, the key question is: how can we model the interpretation of iconic gesture with computational approaches? In the chapter, I first describe the task formally, then propose a computational approach to compute multimodal semantics of utterances. Finally, I describe the experiments which evaluated the approach, and discuss the results.

7.1 Represent multimodal utterances with semantic concepts

Previously, I have discussed multimodal descriptions composed of natural language and iconic gestures, as well as the semantic relations between iconic gestures and the accompanying



Figure 7.1: 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).

speech. Yet, it is still unclear how to computationally derive the semantics of iconic gestures and build corresponding multimodal semantics together with the accompanying verbal content. In this chapter, I 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, 2008b; 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.

The work in this chapter is based on the data from the SAGA corpus (Lücking et al., 2010). Figure 7.1 shows a multimodal utterance from the SAGA corpus. When describing *two lanterns*, a person described “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**.

From the SAGA corpus, I 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, I 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.

7.2 Task formulation

I now describe the task formally. Suppose a verbal utterance U is accompanied by an iconic gesture G (as shown in Figure 7.2), G and U form an ensemble (U, G) that denotes a multimodal utterance. Our ultimate goal is to represent the interpretation of the multimodal utterance

| | |
|-----------------------|--|
| Verbal utterance U | “two, lanterns” |
| Gesture G | <i>two hands facing each other</i> |
| Speech semantics | $[entity, amount]$ |
| Gesture semantics | $[relative\ position, amount]$ |
| Multi-modal semantics | $[entity, relative\ position, amount]$ |

Figure 7.2: Example of a multimodal utterance, and semantic categories.

in a computer system, which can be applied to subsequent tasks such as locating landmarks according to the descriptions (i.e., *referential resolution*). As shown in Figure 7.3, we frame the task of forming such an interpretation into a procedure composed of two phases: map the input information of (U, G) to a set of semantic categories according to the information they convey (as shown in Figure 7.3), then compose the multi-modal semantics of the ensemble with information in the same category across speech and gestures.

We define a function f that computes the semantic types of a speech-gesture ensemble (U, G) . It takes multimodal features of (U, G) as input, and outputs a set of labels c_i which indicates semantic categories. Additionally, we assume that each modality has its own mapping function f_g and f_u , which takes uni-modal features as input, and outputs the semantic categories of corresponding modality (i.e., speech or gesture). In this work, we make the intuitive assumption that the multimodal semantic categories of (U, G) are in fact the union of $f_u(U)$ and $f_g(G)$, as follows:

$$\begin{aligned}
 f_u(U) &= \{c_1, c_2\} \\
 f_g(G) &= \{c_2, c_3\} \\
 f(U, G) &= \{c_1, c_2, c_3\}
 \end{aligned} \tag{7.1}$$

Figure 7.3 illustrates an example of mapping a multimodal utterance “two lanterns G ” to three semantic categories. While the verbal content “two lanterns” are mapped to *amount* and *entity*, the gesture G is mapped to *amount* and *relative positions*. The union of the multimodal semantics are *entity*, *amount* and *relative positions*. The represented interpretation of the ensemble (U, G) is composed of the semantic categories and corresponding values.

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

- **Language features:** The word tokens of each verbal utterance are taken as a bag-of-words to represent linguistic information.

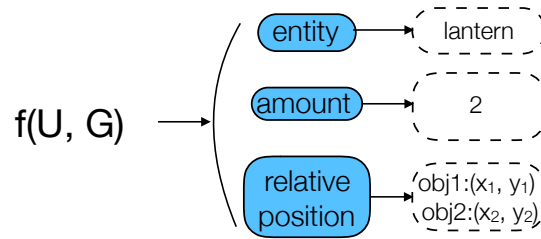


Figure 7.3: Mapping a speech-gesture ensemble to semantic categories in blue rectangles (U and G indicate speech and gesture). Dashed rectangles indicate the value of each semantic category, which are not included in our current work.

- 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).

7.3 Modelling the learning of multimodal semantics

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 \mathbf{X} , we predict a set of semantic category labels $\{c_1, \dots, c_i\}$, 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 machine (SVM) classifier¹ for each semantic label c_i (6 label classifiers in total). Given an input feature \mathbf{X} , we apply all semantic label classifiers to the feature vector. If a semantic label classifier gives positive prediction for input \mathbf{X} , we assign the semantic label to the input. For example, given feature vector of the input utterance “two lanterns”, only the *amount* and *entity* 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.

¹penalty: ℓ_2 , penalty parameter C=1.0, maximum iteration 1000, using an implementation in <http://scikit-learn.org>.

| 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 7.1: Evaluation results. (L and G indicates language and gesture.)

7.4 Experiments

We randomly selected 70% of the gesture-speech ensembles as a training set, using the rest as a test set. Three experiments were designed 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 7.1.

Metrics Following the convention of multi-label classification evaluations, we evaluated our approach with **F1-score**, **accuracy** and **recall** scores.

Gesture features

Since there is no tracked hand motion data, we used the manual annotations to represent gestures. For instance, the gesture in Figure 7.1 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 to 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). There-

fore, gesture feature sequences have variable lengths, in the same sense as utterances have variable amount of word tokens.

7.4.1 Language semantics

As shown in Table 7.1, 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 result in slightly worse performance than language features alone, suggesting that some of the gestural semantics may be complementary to, rather than identical to, the language semantics.

7.4.2 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 that iconic gestures are difficult to interpret without speech (Feyereisen and De Lannoy, 1991). Even humans perform poorly on such a task without verbal content.

In our setup, the abstract gesture features might be one of the reasons of the 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.

7.4.3 Multimodal semantics

As gestures can add meaningful semantic information not present in concurrent speech, we trained and evaluated classifiers on multimodal semantic categories. We assume these are the union of the gesture and language semantics for a given ensemble (as in function f in (7.1) 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 7.1, 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,

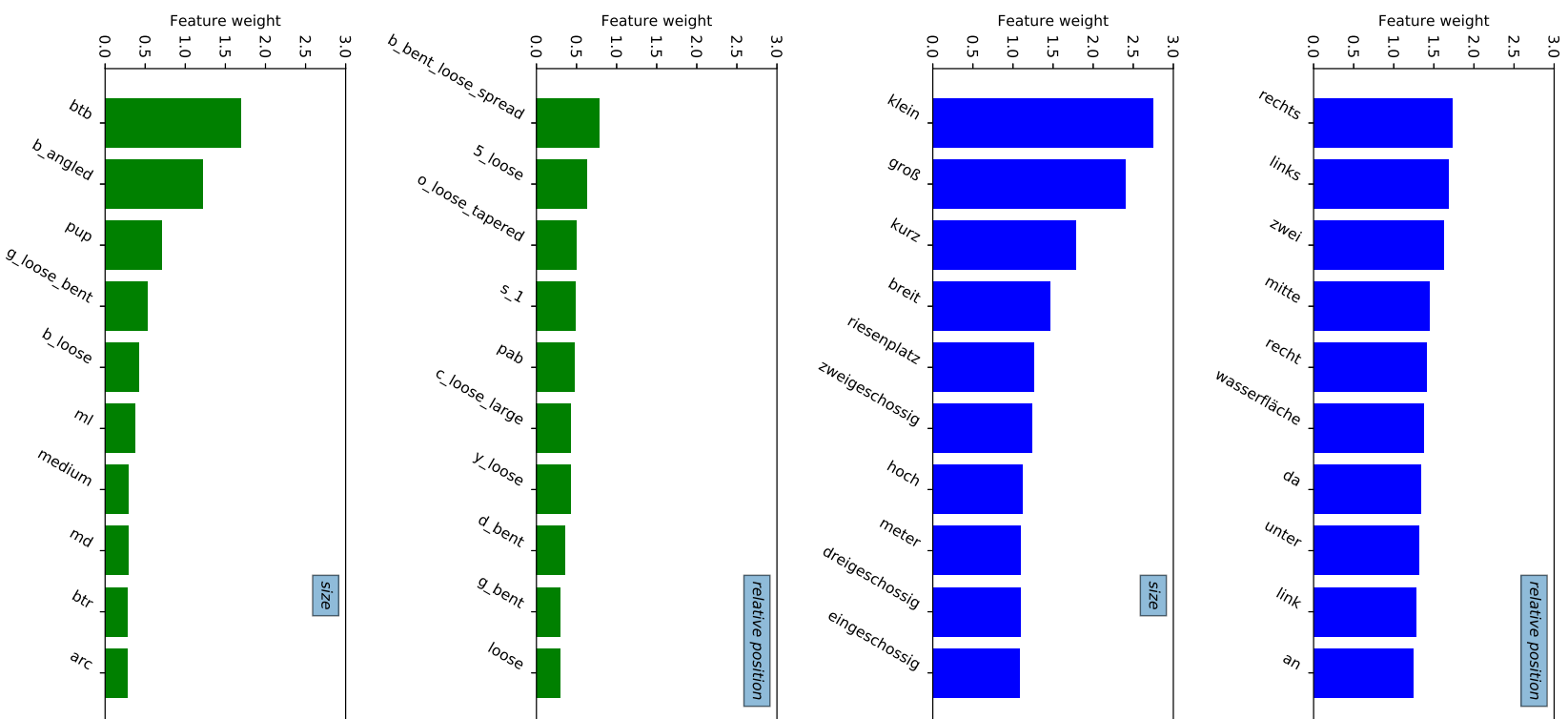


Figure 7.4: Featuring ranking according to coefficient values (weights assigned to the features, see Lücking et al. (2010) for the details of the annotation scheme).

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 7.4 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, such as *b_bent_loose_spread* (an open palm, thumb applied sideways, but not clearly folded and with a weak hand tension) and *5_loose* (an open palm with a weak hand tension) are the two most informative gesture features. 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 feature is back of hand palm direction (*btb*, back of hand palm facing towards body).

7.5 Summary

Language and co-verbal gestures are widely accepted as an integral process of natural communication. In this paper, I 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, I 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. An interesting direction for future work is to predict speech/gesture semantics using raw hand motion features and investigate prediction performance in an online, continuous fashion. This forms part of the ongoing investigation into the interplay of speech and gesture semantics.

8

Conclusion and future work

In this closing chapter, I first give an overview of previous chapters and summarise the works that have been presented, then I conclude this dissertation and discuss the contributions. This is followed by a discussion of future work.

8.1 Overview of the dissertation

Since the seminal work of “Put-that-there” (Bolt, 1998), there has been a large amount of work on multimodal human computer interface, mainly focusing on pointing gestures, symbolic gestures/pen input with pre-defined meanings. These HCIs can only understand a limited set of pre-defined gestures/pen input which convey meanings on their own. However, natural communications often involve co-verbal gestures/pen input whose meanings are related to the accompanying speech. This dissertation aims to work towards a more flexible human-computer interface that can understand multimodal descriptions, where the meaning of hand gestures/pen input (i.e., hand-drawn sketches) relates to accompanied the verbal content, rather than being pre-defined.

I started this dissertation with a look at the background of multimodal communication (Chapter 2). In particular, I focused on multimodal communication composed of natural language and co-verbal hand gestures/sketches, especially on gesture studies of gesture meanings

and the relation between speech and co-verbal gestures/pen input. I also gave an overview of works on multimodal human-computer interfaces (HCIs) which model the interpretation of speech and co-verbal hand gestures. Being well-informed with the knowledge in gesture studies and HCIs, I discussed the limitations of current HCIs and how the knowledge in gesture studies can be deployed to build more general HCIs.

After introducing the background, I presented two multimodal corpora: the *Spatial Scene Description Corpus* and the *Multimodal Object Description Corpus* (Chapter 3). The datasets were collected for works in this dissertation. They are publicly available for researchers in the community.¹ I also briefly introduced the Bielefeld SAGA corpus - a multimodal corpus of route giving and following dialogues (Lücking et al., 2010), which were used to conduct the experiments in Chapter 7.

Chapter 4 presented a system that interprets spatial scene descriptions. While natural language provides size, colour and shape information, abstract deictic gestures denote the positional information and spatial layout of objects. Hence, only when being combined with speech, the abstract deictic gestures get a concrete meaning. After describing the framework of the system, 3 variants of representations were presented, namely, a verbatim representation, pre-defined property label representation and automatic clustering representation. The three representation variants were evaluated with a scene retrieving task. The results show that the automatically performs best, and it overcomes the limitations of Variant A and Variant B. In particular, it reduces the number of symbolic labels need to be kept in the system and automatically learns a set of symbolic labels rather than pre-defined labels, making it easy to scale to larger application domain with more complex objects, such as real-life landmarks.

After exploring the representation methods, I presented a real-time system towards understanding multimodal descriptions incrementally. Using data from the Spatial Description Corpus, an utterance segmenter, NLU module, and a deictic gesture detector were trained for individual system components. The system was evaluated in a real-time manner by replaying the recorded data. The results show that abstract deictic gestures improve the overall accuracy of the interpretation task. Moreover, the deictic gestures also lead to an earlier final correct interpretation decision of the system due to its parallel nature with the verbal content.

While Chapter 4 and 5 focus on abstract deictic gestures, Chapter 6 presented a study that investigates the interplay of symbolic (natural language) and iconic (hand-drawn sketches) modes in multimodal object descriptions. The meaning of symbolic and iconic meanings were modelled with two models originally introduced in previous works and evaluated the contribu-

¹A Spatial Scene Description Corpus https://tingh.github.io/resources/scene_description; Draw and Tell: a Multimodal Object Description Corpus https://tingh.github.io/resources/object_description

tions of natural language and sketches with an image retrieval task. The results show that even adding limited details of sketches improves the performance of the image retrieving model. I also discussed how this relates to the modelling of iconic gesture meanings.

The experiment results also show that iconic information in sketches can be effectively represented with feature vectors. As iconic gestures also convey meaning by resembling visual similarities, It's likely that iconic gestures can also be encoded as vectors in a similar way. Currently, the major challenge lies in a lack of large scale dataset required for training deep learning networks to encode iconic gestures.

Chapter 7 presented an approach that computes semantics of multimodal route giving descriptions. Inspired by Kopp and Bergmann (2017a); Bergmann et al. (2013b) which deploys the coordination between speech and co-verbal hand gestures in a multimodal behaviour generation task, I framed the task of computing semantic categories of iconic gestures as a multi-label prediction problem with words in verbal utterances and hand gesture features as input. The results show that natural language is informative for the interpretation of iconic gestures.

Given the experiments and evaluations described above, I conclude that:

- Due to the parallel nature of speech and gestures, deictics can lead to better and earlier interpretations of multimodal descriptions in an incremental setup, which is essential for triggering immediate clarification requires in dialogue systems;
- Multimodal object descriptions composed of sketches (iconic elements) and verbal utterances outperforms verbal-only or sketch-only descriptions;
- For co-verbal gestures which only receives coherent interpretation together with accompanied verbal content, the verbal content is informative for interpreting such co-verbal gestures.

8.2 Future work

In this dissertation, I have addressed several research questions related to interpreting and applying multimodal descriptions. The presented work also opens up several future research questions that will lead to more general real-time multimodal interfaces. In particular, hereby I discuss following directions for future work:

- **Build large scale multimodal corpora:** A key challenge of modelling the interpretation and application of natural multimodal communication with computational methods (e.g., deep learning neural networks) is to collect large scale datasets for training and evaluating the computational models. To date, despite the availability of portable video/audio recording devices and motion tracking devices, there is still a lack of large scale

multimodal datasets that includes verbal utterances and hand motion data. A potential research direction for future work is to build large scale multimodal datasets. Recently, crowdsourcing platforms such as Amazon Mechanical Turk² and Crowdfunder³ have become more and more popular for collecting large scale datasets, though currently these platforms are most convenient for collecting textual data and sketches. Other approaches for collecting video/audio recordings include retrieving video recordings from Youtube⁴, which provides large amount of data involving real-life conversations. I leave it as future work to explore the most convenient data collection approach.

- **Modelling the interpretation of co-verbal iconic gestures:** Due to a lack of data, in this dissertation, I only investigated the interplay of semantics between symbolic and iconic modes with hand-drawn sketches. One of the future work directions is to model the interpretation of co-verbal iconic gestures. It's an interesting task to investigate whether it's possible to represent iconic gestures with feature vectors in the way we encoded sketch strokes. Encoding iconic gestures into feature vectors will facilitate corresponding multimodal fusion tasks and application tasks such as image retrieval with multimodal descriptions composed of verbal utterances and iconic gestures. As iconic gestures are more abstract and distorted than sketches, it's also challenging to collect enough amount of data for training neural networks that can encode iconic gestures.
- **Build a general real-time multimodal system:** The ultimate goal of learning to interpret and apply multimodal descriptions is to build a general multimodal human-computer interface that can process what comes natural. While in this dissertation I have made a first effort to learn the semantics of co-verbal gestures using natural language information and hand gesture features, due to a lack of data, it was not possible to build computational models to extract and represent iconic information from raw hand gestures. I leave it as future work to build a general real-time system that can learn to extract information from co-verbal gestures, represent the information in the system and apply the represented knowledge to a real-life task such as image retrieval or navigation tasks.

²<https://www.mturk.com/>

³<http://www.crowdfunder.com/>

⁴www.youtube.com/

Bibliography

- Harika Abburi, Rajendra Prasath, Manish Shrivastava, and Suryakanth V Gangashetty. Multimodal sentiment analysis using deep neural networks. In *International Conference on Mining Intelligence and Knowledge Exploration*, pages 58–65. Springer, 2016.
- Martha W Alibali, Sotaro Kita, and Amanda J Young. Gesture and the process of speech production: We think, therefore we gesture. *Language and cognitive processes*, 15(6):593–613, 2000.
- Mw Alibali. Gesture in spatial cognition: Expressing, communicating, and thinking about spatial information. *Spatial Cognition and Computation*, 5(4):307–331, 2005.
- Yoav Artzi and Luke Zettlemoyer. Weakly supervised learning of semantic parsers for mapping instructions to actions. *Transactions of the Association for Computational Linguistics*, 1:49–62, 2013.
- Nicholas Asher and Alex Lascarides. *Logics of conversation*. Cambridge University Press, 2003.
- Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379, 2010.
- Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.
- Timo Baumann and David Schlangen. The inprotk 2012 release. In *NAACL-HLT Workshop on Future Directions and Needs in the Spoken Dialog Community: Tools and Data*, pages 29–32. Association for Computational Linguistics, 2012.
- Mark Bayazit, Alex Couture-Beil, and Greg Mori. Real-time motion-based gesture recognition using the gpu. In *Conference on Machine Vision Applications*, pages 9–12, 2009.

- Geoffrey Beattie and Heather Shovelton. Mapping the range of information contained in the iconic hand gestures that accompany spontaneous speech. *Journal of language and social psychology*, 18(4):438–462, 1999.
- Azzedine Bendjebbour, Yves Delignon, Laurent Fouque, Vincent Samson, and Wojciech Pieczynski. Multisensor image segmentation using dempster-shafer fusion in markov fields context. *IEEE Transactions on Geoscience and Remote Sensing*, 39(8):1789–1798, 2001.
- Kirsten Bergmann and Stefan Kopp. Multimodal content representation for speech and gesture production. In *Proceedings of the 2nd Workshop on Multimodal Output Generation*, pages 61–68, 2008a.
- Kirsten Bergmann and Stefan Kopp. Multimodal content representation for speech and gesture production. In *Proceedings of the 2nd Workshop on Multimodal Output Generation*, pages 61–68, 2008b.
- Kirsten Bergmann, Volkan Aksu, and Stefan Kopp. 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)*, 2011.
- Kirsten Bergmann, Florian Hahn, Stefan Kopp, Hannes Rieser, and Insa Röpke. Integrating gesture meaning and verbal meaning for german verbs of motion: Theory and simulation. In *Proceedings of the Tilburg Gesture Research Meeting (TiGeR 2013)*, 2013a.
- Kirsten Bergmann, Sebastian Kahl, and Stefan Kopp. Modeling the semantic coordination of speech and gesture under cognitive and linguistic constraints. In *International Workshop on Intelligent Virtual Agents*, pages 203–216. Springer, 2013b.
- Kirsten Bergmann, Sebastian Kahl, and Stefan Kopp. How is information distributed across speech and gesture? a cognitive modeling approach. *Cognitive Processing, Special Issue: Proceedings of KogWis*, pages S84–S87, 2014.
- Philippe Blache, Roxane Bertrand, and Gaëlle Ferré. Creating and exploiting multimodal annotated corpora: the toma project. In *Multimodal corpora*, pages 38–53. Springer, 2009.
- Richard A Bolt. “put-that-there”: Voice and gesture at the graphics interface. *ACM Siggraph Computer Graphics*, 32(4), 1998.
- Tad T Brunyé and Holly A Taylor. Working memory in developing and applying mental models from spatial descriptions. *Journal of Memory and Language*, 58(3):701–729, 2008.

- Ellen Campana, Laura Silverman, Michael K Tanenhaus, Loisa Bennetto, and Stephanie Packard. Real-time integration of gesture and speech during reference resolution. In *Proceedings of the 27th annual meeting of the Cognitive Science Society*, pages 378–383. Cite-seer, 2005.
- Justine Cassell, Timothy Bickmore, Mark Billingham, Lee Campbell, Kenny Chang, Hannes Vilhjálmsón, and Hao Yan. Embodiment in conversational interfaces: Rea. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 520–527. Association for Computing Machinery, 1999.
- Justine Cassell, Stefan Kopp, Paul Tepper, Kim Ferriman, and Kristina Striegnitz. Trading spaces: How humans and humanoids use speech and gesture to give directions. *Conversational informatics*, pages 133–160, 2007.
- Joyce Chai, Shimei Pan, Michelle X Zhou, and Keith Houck. Context-based multimodal input understanding in conversational systems. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, pages 87–92. IEEE, 2002.
- Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 163–171. Association for Computing Machinery, 2017.
- Hong Cheng, Lu Yang, and Zicheng Liu. Survey on 3d hand gesture recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(9):1659–1673, 2016.
- François Chollet. Keras. <https://github.com/fchollet/keras>, 2015.
- Kawai Chui. Temporal patterning of speech and iconic gestures in conversational discourse. *Journal of Pragmatics*, 37(6):871–887, 2005.
- R B Church and S Goldin-Meadow. The mismatch between gesture and speech as an index of transitional knowledge. *Cognition*, 23(1):43–71, 1986.
- Philip R Cohen, Michael Johnston, David McGee, Sharon Oviatt, Jay Pittman, Ira Smith, Liang Chen, and Josh Clow. Quickset: Multimodal interaction for distributed applications. In *Proceedings of the fifth ACM international conference on Multimedia*, pages 31–40. Association for Computing Machinery, 1997.
- Guillem Collell Talleda and Marie-Francine Moens. Is an image worth more than a thousand words? on the fine-grain semantic differences between visual and linguistic representations.

- In *Proceedings of the 26th International Conference on Computational Linguistics*. The association of computational linguistics, 2016.
- Bruno Dumas, Denis Lalanne, and Sharon Oviatt. Multimodal interfaces: A survey of principles, models and frameworks. In *Human machine interaction*, pages 3–26. Springer, 2009.
- Mathias Eitz, Kristian Hildebrand, Tamy Boubekeur, and Marc Alexa. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. *IEEE transactions on visualization and computer graphics*, 17(11):1624–1636, 2011.
- Karen Emmorey, Barbara Tversky, and Holly a. Taylor. Using space to describe space: Perspective in speech, sign, and gesture. *Spatial Cognition and Computation*, 2(3):157–180, 2000a.
- Karen Emmorey, Barbara Tversky, and Holly a. Taylor. Using space to describe space: Perspective in speech, sign, and gesture. *Spatial Cognition and Computation*, 2(3):157–180, 2000b.
- Julien Epps, Sharon Oviatt, and Fang Chen. Integration of speech and gesture inputs during multimodal interaction. In *Proceedings of the Australian International Conference on Computer-Human Interaction (OzCHI)*, 2004.
- Pierre Feyereisen and Jacques-Dominique De Lannoy. *Gestures and speech: Psychological investigations*. Cambridge University Press, 1991.
- Stavroula-Evita Fotinea, Eleni Efthimiou, Maria Koutsombogera, Athanasia-Lida Dimou, Theodore Goulas, and Kyriaki Vasilaki. Multimodal resources for human-robot communication modelling. In *Language Resources and Evaluation Conference*, 2016.
- William T Freeman and Michal Roth. Orientation histograms for hand gesture recognition. In *International workshop on automatic face and gesture recognition*, volume 12, pages 296–301, 1995.
- Farina Freigang and Stefan Kopp. This is what’s important—using speech and gesture to create focus in multimodal utterance. In *International Conference on Intelligent Virtual Agents*, pages 96–109. Springer, 2016.
- Farina Freigang, Sören Klett, and Stefan Kopp. Pragmatic multimodality: Effects of nonverbal cues of focus and certainty in a virtual human. In *International Conference on Intelligent Virtual Agents*, pages 142–155. Springer, 2017.

- Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. In *9th International Conference on Artificial Neural Networks*. IET, 1999.
- Gianluca Giorgolo. *Space and Time in Our Hands*. PhD thesis, Netherlands Graduate School of Linguistics, 2010.
- Susan Goldin-Meadow. *Hearing gesture: How our hands help us think*. Harvard University Press, 2005.
- Peter Gorniak and Deb Roy. Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, 21:429–470, 2004.
- Peter Gorniak and Deb Roy. Probabilistic Grounding of Situated Speech using Plan Recognition and Reference Resolution. In *In Proceedings of the Seventh International Conference on Multimodal Interfaces (ICMI 2005)*, pages 138–143, 2005.
- Kalanit Grill-Spector and Nancy Kanwisher. Visual recognition as soon as you know it is there, you know what it is. *Psychological Science*, 16(2):152–160, 2005.
- Ting Han, Casey Kennington, and David Schlangen. Building and Applying Perceptually-Grounded Representations of Multimodal Scene Descriptions. In *Proceedings of the 19th SemDial Workshop on the Semantics and Pragmatics of Dialogue (goDIAL)*, 2015.
- Ting Han, Casey Kennington, and David Schlangen. 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, 2018.
- Stevan Harnad. The Symbol Grounding Problem. *Physica D: Nonlinear Phenomena*, 42: 335–346, 1990.
- Haitham Hasan and S Abdul-Kareem. Retracted article: Static hand gesture recognition using neural networks. *Artificial Intelligence Review*, 41(2):147–181, 2014.
- Lode Hoste, Bruno Dumas, and Beat Signer. Mudra: a unified multimodal interaction framework. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 97–104. Association for Computing Machinery, 2011.
- Autumn B Hostetter, Martha W Alibali, and Sotaro Kita. I see it in my hands’ eye: Representational gestures reflect conceptual demands. *Language and Cognitive Processes*, 22(3): 313–336, 2007.

- Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4555–4564, 2016.
- Pui-Yu Hui and Helen Meng. Latent Semantic Analysis for Multimodal User Input With Speech and Gestures. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(2):417–429, February 2014.
- Alexandra Jesse and Elizabeth K Johnson. Prosodic temporal alignment of co-speech gestures to speech facilitates referent resolution. *Journal of Experimental Psychology: Human Perception and Performance*, 38(6):1567, 2012.
- Michael Johnston, Philip R Cohen, David Mcgee, Sharon L Oviatt, James A Pittman, and Ira Smith. Unification-based Multimodal Integration. In *In Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics*, pages 281–288, Madrid, 1997.
- Michael Johnston, Srinivas Bangalore, Gunaranjan Vasireddy, Amanda Stent, Patrick Ehlen, Marilyn Walker, Steve Whittaker, and Preetam Maloor. MATCH: An architecture for multimodal dialogue systems. *Computational Linguistics*, pages 376–383, 2002.
- Chris Joslin, Ayman El-Sawah, Qing Chen, and Nicolas Georganas. Dynamic gesture recognition. In *Instrumentation and Measurement Technology Conference, 2005. IMTC 2005. Proceedings of the IEEE*, volume 3, pages 1706–1711. IEEE, 2005.
- Hans Kamp and Uwe Reyle. *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory*, volume 42. Springer Science & Business Media, 2013.
- Maria Karam and M. C. Schraefel. A Taxonomy of Gestures in Human Computer Interactions. Technical report, University of Southampton, 2005.
- Adam Kendon. Gesticulation and speech: two aspects of the process of utterance. *The Relationship of Verbal and Nonverbal Communication*, 25:207–227, 1980a.
- Adam Kendon. Gesticulation and speech: Two aspects of the process of utterance. *The relationship of verbal and nonverbal communication*, 25(1980):207–227, 1980b.
- Adam Kendon. Gesture. *Annual Review of Anthropology*, 26(1):109–128, 1997.

- Casey Kennington and David Schlangen. Simple Learning and Compositional Application of Perceptually Grounded Word Meanings for Incremental Reference Resolution. In *Proceedings of the association of computational linguistics*, Beijing, China, 2015. Association for Computational Linguistics.
- Casey Kennington, Spyros Kousidis, and David Schlangen. Interpreting Situated Dialogue Utterances: an Update Model that Uses Speech, Gaze, and Gesture Information. In *Special interest group on discourse and dialogue 2013*, 2013.
- Casey Kennington, Livia Dia, and David Schlangen. A Discriminative Model for Perceptually-Grounded Incremental Reference Resolution. In *Proceedings of IWCS*. Association for Computational Linguistics, 2015.
- Walter Kintsch and Teun a. van Dijk. Toward a model of text comprehension and production. *Psychological Review*, 85(5):363–394, 1978.
- Sotaro Kita. How representational gestures help speaking. *Language and gesture*, 1, 2000.
- Sotaro Kita and Asli Özyürek. 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, 2003.
- Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, volume 2, 2015.
- Thomas Kollar, Stefanie Tellex, Deb Roy, and Nicholas Roy. Toward understanding natural language directions. *Proceeding of the 5th ACMIEEE international conference on Human-robot interaction HRI 10*, page 259, 2010.
- David B Koons, Carlton J Sparrell, and Kristinn R Thorisson. Integrating simultaneous input from speech, gaze and hand gestures. In Mark T. Maybury, editor, *Intelligent Multimedia Interfaces*, pages 257–276. AAAI Press/ MIT Press, Cambridge, MA, 1993a.
- David B Koons, Carlton J Sparrell, and Kristinn Rr Thorisson. Integrating simultaneous input from speech, gaze, and hand gestures. *MIT Press: Menlo Park, CA*, pages 257–276, 1993b.
- Stefan Kopp and Kirsten Bergmann. Using cognitive models to understand multimodal processes: The case for speech and gesture production. In Sharon Oviatt, Björn Schuller, Phil Cohen, and Antonio Krüger, editors, *Handbook of Multimodal-Multisensor Interfaces*. ACM Books, Morgan Claypool, 2017a.

- Stefan Kopp and Kirsten Bergmann. Using Cognitive Models to Understand Multimodal Processes: The Case for Speech and Gesture Production. In Sharon Oviatt, Björn Schuller, Phil Cohen, and Antonio Krüger, editors, *Handbook of Multimodal-Multisensor Interfaces*. ACM Books, Morgan Claypool, 2017b.
- Stefan Kopp, Paul Tepper, and Justine Cassell. Towards integrated microplanning of language and iconic gesture for multimodal output. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 97–104. Association for Computing Machinery, 2004.
- Spyridon Kousidis, Thies Pfeiffer, Zofia Malisz, Petra Wagner, and David Schlangen. Evaluating a minimally invasive laboratory architecture for recording multimodal conversational data. In *Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog, INTERSPEECH2012 Satellite Workshop*, pages 39–42, 2012.
- Spyros Kousidis, Casey Kennington, and David Schlangen. Investigating speaker gaze and pointing behaviour in human-computer interaction with the mint. tools collection. In *Proceedings of the Special interest group on discourse and dialogue 2013*, pages 319–323, 2013a.
- Spyros Kousidis, Thies Pfeiffer, and David Schlangen. MINT . tools : Tools and Adaptors Supporting Acquisition , Annotation and Analysis of Multimodal Corpora. In *Proceedings of Interspeech 2013*, pages 2649–2653, Lyon, France, 2013b. ISCA.
- S Larsson. Formal semantics for perceptual classification. *Journal of Logic and Computation*, 2013.
- Alex Lascarides and Matthew Stone. A formal semantic analysis of gesture. *Journal of Semantics*, 26(4):393–449, 2009.
- Thomas Leonard and Fred Cummins. Temporal alignment of gesture and speech. *Proceedings of Gespın*, pages 1–6, 2009.
- Michael Levit and Deb Roy. Interpretation of spatial language in a map navigation task. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 37(3):667–679, 2007.
- Bo Li, Tobias Schreck, Afzal Godil, Marc Alexa, Tamy Boubekeur, Benjamin Bustos, Jipeng Chen, Mathias Eitz, Takahiko Furuya, Kristian Hildebrand, et al. Shrec’12 track: Sketch-based 3d shape retrieval. In *Eurographics Workshop on 3D Object Retrieval*, pages 109–118, 2012.

- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.
- Li Liu and Ling Shao. Learning discriminative representations from rgb-d video data. In *International Joint Conferences on Artificial Intelligence*, volume 4, page 8, 2013.
- Edward Loper and Steven Bird. NLTK: The natural language toolkit. In *Proceedings of ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*, pages 63–70. Association for Computational Linguistics, 2002.
- Lorenzo Lucignano, Francesco Cutugno, Silvia Rossi, and Alberto Finzi. A dialogue system for multimodal human-robot interaction. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 197–204. Association for Computing Machinery, 2013.
- Andy Lücking, Kirsten Bergmann, Florian Hahn, Stefan Kopp, and Hannes Rieser. The bielefeld speech and gesture alignment corpus (saga). In *The workshop of Multimodal corpora—advances in capturing, coding and analyzing multimodality in Language Resources and Evaluation Conference*, 2010.
- Giulio Marin, Fabio Dominio, and Pietro Zanuttigh. Hand gesture recognition with leap motion and kinect devices. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 1565–1569. IEEE, 2014.
- Cynthia Matuszek, Liefeng Bo, Luke Zettlemoyer, and Dieter Fox. Learning from Unscripted Deictic Gesture and Language for Human-Robot Interactions. In *The Proceedings of Association for the Advancement of Artificial Intelligence 2014*, 2014.
- P. McGuire, J. Fritsch, J.J. Steil, F. Rothling, G.a. Fink, S. Wachsmuth, G. Sagerer, and H. Ritter. Multi-modal human-machine communication for instructing robot grasping tasks. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 2, page 7, 2002.
- David McNeill. *Hand and Mind: What Gestures Reveal About Thought*. University of Chicago press, 1992.
- David McNeill. *Gesture and Thought*, volume 18. University of Chicago Press, 2005.

- David McNeill. Gesture: a psycholinguistic approach. *The encyclopedia of language and linguistics*, pages 58–66, 2006.
- David McNeill and Susan D Duncan. Growth points in thinking-for-speaking. *Language and Gestures*, pages 141–161, 1998.
- David McNeill, Justine Cassel, and Elena T. Levy. Abstract deixis. *Semiotica*, 95(1-2):5–20, 1993.
- Ian Moar and Gordon H Bower. Inconsistency in spatial knowledge. *Memory & Cognition*, 11(2):107–113, 1983.
- Pavlo Molchanov, Shalini Gupta, Kihwan Kim, and Jan Kautz. Hand gesture recognition with 3d convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1–7, 2015.
- GRS Murthy and RS Jadon. A review of vision based hand gestures recognition. *International Journal of Information Technology and Knowledge Management*, 2(2):405–410, 2009.
- Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2016.
- Kai Nickel and Rainer Stiefelhagen. Pointing Gesture Recognition based on 3D-Tracking of Face , Hands and Head Orientation Categories and Subject Descriptors. *Proceedings of the 5th international conference on Multimodal interfaces*, pages 140–146, 2003.
- John Niekrasz and Matthew Purver. A multimodal discourse ontology for meeting understanding. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 3869 LNCS, pages 162–173, 2006.
- Mark Nieuwenstein and Brad Wyble. Beyond a mask and against the bottleneck: Retroactive dual-task interference during working memory consolidation of a masked visual target. *Journal of Experimental Psychology: General*, 143(3):1409, 2014.
- Shuichi Nobe. Where do most spontaneous representational gestures actually occur with respect to speech. *Language and Gesture*, 2:186, 2000.
- Sharon Oviatt. Multimodal interfaces. *The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications*, 14:286–304, 2003.

- Sharon Oviatt and Philip Cohen. Perceptual user interfaces: multimodal interfaces that process what comes naturally. *Communications of the Association for Computing Machinery*, 43(3): 45–53, 2000.
- Aslı Özyürek, Roel M Willems, Sotaro Kita, and Peter Hagoort. On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials. *Journal of cognitive neuroscience*, 19(4):605–616, 2007.
- Charles Sanders Pierce. On a new list of categories. In Charles Hartshorne and Paul Weiss, editors, *C.S. Pierce: The Collected Papers*. Harvard University Press, Cambridge, M.A., USA, 1867.
- Karen J Pine, Nicola Lufkin, Elizabeth Kirk, and David Messer. A microgenetic analysis of the relationship between speech and gesture in children: Evidence for semantic and temporal asynchrony. *Language and Cognitive Processes*, 22(2):234–246, 2007.
- Vassilis Pitsikalis, Athanasios Katsamanis, Stavros Theodorakis, and Petros Maragos. Multi-modal gesture recognition via multiple hypotheses rescoring. In *Gesture Recognition*, pages 467–496. Springer, 2017.
- Massimo Poesio and David R Traum. Conversational actions and discourse situations. *Computational intelligence*, 13(3):309–347, 1997.
- Xueming Qian, Xianglong Tan, Yuting Zhang, Richang Hong, and Meng Wang. Enhancing sketch-based image retrieval by re-ranking and relevance feedback. *IEEE Transactions on Image Processing*, 25(1):195–208, 2016.
- Francis Quek, David McNeill, Robert Bryll, Susan Duncan, Xin-Feng Ma, Cemil Kirbas, Karl E McCullough, and Rashid Ansari. Multimodal human discourse: gesture and speech. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 9(3):171–193, 2002.
- J Donald Ragsdale and Catherine Fry Silvia. Distribution of kinesic hesitation phenomena in spontaneous speech. *Language and Speech*, 25(2):185–190, 1982.
- Siddharth S Rautaray and Anupam Agrawal. Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review*, 43(1):1–54, 2015.
- Hilke Reckman, Jeff Orkin, and Deb Roy. Learning meanings of words and constructions, grounded in a virtual game. *Semantic Approaches in Natural Language Processing*, page 67, 2010.

- Hannes Rieser. When hands talk to mouth. gesture and speech as autonomous communicating processes. In *Proceedings of the 19th workshop on the semantics and pragmatics of dialogue*, page 122, 2015.
- Hannes Rieser and Massimo Poesio. Interactive gesture in dialogue: A ptt model. In *Proceedings of the 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 87–96. Association for Computational Linguistics, 2009.
- Wolff-Michael Roth. From action to discourse: The bridging function of gestures. *Cognitive Systems Research*, 3(3):535–554, 2002.
- Deb Roy and Ehud Reiter. Connecting language to the world. *Artificial Intelligence*, 167(1-2): 1–12, 2005.
- Deb K Roy. Learning visually grounded words and syntax for a scene description task. *Computer speech & language*, 16(3):353–385, 2002.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.
- Amir Sadeghipour and Stefan Kopp. Learning a motor grammar of iconic gestures. In *Proceedings of the 35th annual meeting of the Cognitive Science Society (CogSci 2013)*, 2014.
- Amir Sadeghipour and Louis-Philippe Morency. 3D Iconic Gesture Dataset, 2011.
- Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35(4):119, 2016.
- Emanuel A Schegloff. On some gestures’ relation to talk. *Structures of social action: Studies in conversation analysis*, pages 266–296, 1984.
- Florian Schiel, Silke Steininger, and Ulrich Türk. The smartkom multimodal corpus at bas. In *Language Resources and Evaluation Conference*, 2002.
- David Schlangen and Gabriel Skantze. A general, abstract model of incremental dialogue processing. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 710–718. Association for Computational Linguistics, 2009.

- David Schlangen, Sina Zarrieß, and Casey Kennington. Resolving references to objects in photographs using the words-as-classifiers model. In *Proceedings of the association of computational linguistics 2016*, Berlin, Germany, August 2016.
- Laura F Schneider and Holly a. Taylor. How do you get there from here? Mental representations of route descriptions. *Applied Cognitive Psychology*, 13(September 1998):415–441, 1999.
- Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.
- Björn Schuller, Ronald Müller, Manfred Lang, and Gerhard Rigoll. Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In *Ninth European Conference on Speech Communication and Technology*, 2005.
- Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, volume 2, 2015.
- K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.
- Cees GM Snoek, Marcel Worring, and Arnold WM Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402. Association for Computing Machinery, 2005.
- Timo Sowa. *Understanding Coverbal Ionic Gestures in Shape Descriptions*. IOS Press Amsterdam, 2006.
- Timo Sowa and Stefan Kopp. A cognitive model for the representation and processing of shape-related gestures. In *Proceedings of the European Cognitive Science Conference (EuroCogSci03)*, page 441. Citeseer, 2003.
- Timo Sowa and Ipke Wachsmuth. Coverbal iconic gestures for object descriptions in virtual environments: An empirical study. *Gestures. Meaning and Use*, pages 365–376, 2003.
- Timo Sowa and Ipke Wachsmuth. A model for the representation and processing of shape in coverbal iconic gestures. In *Proceeding of KogWis 05: the German Cognitive Science Conference*, 2005.

- Timo Sowa and Ipke Wachsmuth. A computational model for the representation and processing of shape in coverbal iconic gestures. In *Spatial Language and Dialogue*. Oxford : Oxford University Press, 2009.
- Luc Steels and Frederic Kaplan. Situated grounded word semantics. In *International Joint Conference on Artificial Intelligence*, volume 2, pages 862–867, 1999.
- R. Stiefelhagen, C. Fugen, R. Gieselmann, H. Holzapfel, K. Nickel, and A. Waibel. 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, 2004.
- K Striegnitz, P Tepper, A Lovett, and J Cassell. Knowledge Representation for Generating Locating Gestures in Route Directions. *Spatial Language in Dialogue*, 1:1–13, 2005.
- Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics*, 33, August 2014.
- Grigorios Tsoumakos and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3), 2006.
- Matthew Turk. Multimodal interaction: A review. *Pattern Recognition Letters*, 36:189–195, 2014.
- Barbara Tversky, Julie Heiser, Paul Lee, and Marie-Paule Daniel. Explanations in Gesture, Diagram, and Word. In Kenny R. Coventry, Thora Tenbrink, and John Bateman, editors, *Spatial Language and Dialogue*, pages 119–131. Oxford University Press, apr 2009.
- Adam Vogel and Dan Jurafsky. Learning to Follow Navigational Directions. In *Proceedings of the association of computational linguistics*, pages 806–814, 2010.
- David Whitney, Miles Eldon, John Oberlin, and Stefanie Tellex. Interpreting multimodal referring expressions in real time. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pages 3331–3338. IEEE, 2016.
- Di Wu, Lionel Pigou, Pieter-Jan Kindermans, Nam Do-Hoang Le, Ling Shao, Joni Dambre, and Jean-Marc Odobez. Deep dynamic neural networks for multimodal gesture segmentation and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 38(8): 1583–1597, 2016.

- Jonathan Wu, Janusz Konrad, and Prakash Ishwar. Dynamic time warping for gesture-based user identification and authentication with kinect. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 2371–2375. IEEE, 2013.
- Lizhong Wu, Sharon L. Oviatt, and Philip R. Cohen. Multimodal integration—a statistical view. *IEEE Transactions on Multimedia*, 1(4):334–341, 1999.
- Ying Wu and Thomas S Huang. Vision-based gesture recognition: A review. In *International Gesture Workshop*, pages 103–115. Springer, 1999.
- Jianxiong Xiao, Krista A Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Sun database: Exploring a large collection of scene categories. *International Journal of Computer Vision*, 119(1):3–22, 2016.
- Huaxin Xu and Tat-Seng Chua. Fusion of av features and external information sources for event detection in team sports video. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2(1):44–67, 2006.
- Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 799–807, 2016.
- Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention recurrent network for human communication comprehension. *arXiv preprint arXiv:1802.00923*, 2018.
- Yuanxin Zhu, Guangyou Xu, and David J Kriegman. A Real-Time Approach to the Spotting, Representation, and Recognition of Hand Gestures for Human–Computer Interaction. *Computer Vision and Image Understanding*, 85(3):189–208, 2002.