Learning to Describe Multimodally from Parallel Unimodal Data?
A Pilot Study on Verbal and Sketched Object Descriptions

Ting Han[1], Sina Zarrieß[1], Kazunori Komatani[2], David Schlangen[1]
[1]Dialogue Systems Group/Bielefeld University
[1]firstname.lastname@uni-bielefeld.de
[2]The Institute of Scientific and Industrial Research, Osaka University
[2]komatani@sanken.osaka-u.ac.jp

Abstract

Previous work on multimodality in interaction has mostly focussed on integrating models for verbal utterances and embodied modalities like gestures. In this paper, we take a first step towards investigating multimodal interaction that combines verbal utterances and hand-drawn sketches which can be essential, for instance, for conveying explanation in dialogue. While there is a lot of theoretical work on how drawing and sketching convey iconic meaning, there is no realistic data set that pairs language and sketch as integrated modalities. Recently, the *Draw-and-Tell* corpus enriched a pre-existing dataset (the "Sketchy Dataset") with verbal descriptions of the sketched images. We base our study on this corpus and implement two models that learn to generate simple verbal and sketched object descriptions in a parallel fashion. We evaluated our models in unimodal and multimodal object identification tasks with human listeners via crowd-sourcing experiments. The results show that partial hand-drawn sketches clearly improve the effectiveness of verbal descriptions, even if the generator did not coordinate their meanings. Interestingly, we also find that unimodal sketched object descriptions outperform multimodal descriptions. We argue that this highlights the great potential of sketched explanations for multimodal interaction, but at the same time, shows the need for more natural data sets that provide insights into the orchestration of verbal and sketched elements in multimodal descriptions.

1 Introduction

Human communications are multimodal in nature, in various ways and settings. Research on multi-modality in linguistics, NLP and HRI has often focussed on *embodied* interaction, and studied the complex interplay between speech, gestures, facial expressions, gaze, etc., cf. (McNeill, 1992; Cassell et al., 1994; Kopp et al., 2008; Fang et al., 2015; Gatt and Paggio, 2014; De Ruiter et al., 2012). In other areas, there is a long-standing tradition of looking at other non-verbal modalities (e.g. sketches, paintings, diagrams) as well, as they perfectly illustrate the human capacity of orchestrating various means of expression for abstracting from states of affairs in the real world and convey meaning (DeCarlo and Santella, 2002; Kenneth et al., 2011; Tversky, 2014). Sketches, as a visual modality, naturally occur in multimodal dialogue, for instance in contexts where speakers need to communicate complex concepts or ideas. Sketches are frequently and systematically used by designers, engineers, teachers and students when they need to explain their ideas in interaction (Oltmans and Davis, 2001; Prain and Waldrip, 2006; Adler and Davis, 2007; Tversky et al., 2009; Wetzel and Forbus, 2010).

Whereas the fields of NLP and HRI have come up with methods for studying gestures in multimodal interaction empirically such as (Stiefelhagen et al., 2004; de Wit et al., 2018) and small scale multimodal data collections (Lücking et al., 2010), to the best of our knowledge, there is no dataset of human interactions via verbal utterances and sketches. Han and Schlangen (2017) presented a *Draw-and-Tell* corpus. The corpus augmented an existing corpus, the Sketchy dataset (Sangkloy et al., 2016) that pairs photos with hand-drawn sketches, with verbal object descriptions, providing parallel uni-modal data of photo descriptions. In this paper, we used the *Draw-and-Tell* corpus to explore how to generate multimodal object descriptions, even though sketches and utterances were collected independently in the original corpus.

Original description	Generated Description	Candidate photos
brown with a little white, steeple on top, 3 windows, steep roof	*grey roof*	

Figure 1: Multimodal descriptions of a **church**, in a context with other churches as distractors, target referent is the second from left. Column 1 shows the original human, column 2 the generated description.

We investigate multi-modal descriptions of objects in real-world images using sketches as an additional modality. The description of visual entities in real-world images poses considerable communicative challenges to machines (Zarrieß and Schlangen, 2016), and might be compared to the description of complex objects in the design domain (Adler and Davis, 2007). As an example, the verbal referring expression (RE) in Figure 1 mentions the colour property, whereas the strokes indicate the orientation and shape of the church in the image which is very difficult to express verbally.

In current work, we trained two standard captioning models to generate verbal descriptions for objects in real-life photos in the *Draw-and-Tell* corpus and combine these models with a simple stroke selection approach that represents the target object with reduced iconic information. We evaluated the generated unimodal and multimodal description in an image identification task with humans. As shown in Figure 1, given the original description of the photo that contains a church, we generate a multimodal description, which helped listeners to identify the target photo from a set of candidate photos. We observed some interesting interdependencies between the effectiveness of iconic elements and the underlying generation model: the multimodal descriptions are more effective when the verbal expression is shorter and potentially more ambiguous, while less contradictory with the iconic modality. It is interesting that sketches alone are more effective than when combined with verbal descriptions. While this doesn't contradict the fact that multimodal descriptions are more effective than verbal descriptions, it highlights the great potential of sketches in multimodal interactions, and shows that natural data sets are needed for investigating the orchestration of verbal and sketched elements in multimodal descriptions.

Our contributions are summarised as follows: **1)** We investigate a new task, generating multimodal object descriptions composed of natural language and sketches, which is useful for multimodal explanation in dialogue; **2)** We implement and evaluate two pilot systems for multimodal object description that generate the verbal phrases and select strokes from a sketch in parallel; **3)** We show that even partial sketches with limited visual detail can complement verbal descriptions successfully and are very effective in unimodal conditions as well.

2 Related work

Our work is inspired by recent trends in language & vision, and generally targets the study of multi-modal interaction between humans and machines.

Sketch generation from real-world photographs is a well-known computer vision task that has been worked on for at least 20 years and is also referred to as Non-Photorealistic Rendering (NPR) (Gooch and Gooch, 2001; DeCarlo and Santella, 2002; Tresset and Leymarie, 2013; Ha and Eck, 2018). NPR in particular goes beyond simple edge detection (Canny, 1987) and aims at interpreting an image such that important aspects or causal relations can be depicted in a salient way (DeCarlo and Santella, 2002). Recently, based on the *Quick, Draw!* dataset, Ha and Eck (2018) have presented a neural model that learns to draw sketches of objects like cats from unfinished human-sketches, but not from real-world images. The task of generating human-like sketches from images is still under-explored.

Verbal object and image description generation has received increasing attention in the last years, and is now adressed in a range of sub-tasks in the language & vision community, e.g. for image and scene descriptions (Karpathy and Fei-Fei, 2015; Vinyals et al., 2015), referring expression generation (Mao et al., 2016; Yu et al., 2017), justification generation (Hendricks et al., 2016), and it is also closely related to more interactive settings such as (De Vries et al., 2017). Among these types of object description, referring expressions are probably the most well-known as a linguistic phenomenon in research on situated interaction, and have been studied in depth in the field of NLG and referring expression generation (REG) in particular, cf. (Dale and Reiter, 1995; Krahmer and Van Deemter, 2012). Here, the task is to generate a discriminative, pragmatically appropriate expression that helps a listener to identify a target referent. Our work sits in between classical REG that aims at generating human-like discriminative expressions and image descriptions or explanations, and builds on the descriptions collected by Han and Schlangen (2017). In their setup, participants were prompted to produce attribute centric descriptions by enumerating the properties of a target object as compared to visually similar distractor objects of the same category (e.g. *"brown with a little white, steeple on top, 3 windows, steep roof"* in Figure 1). We believe this is an interesting starting point for investigating into complex, multi-modal object descriptions and explanations, which is more feasible than real-world scenarios such as interfaces for engineers or designers (Adler and Davis, 2007; Wetzel and Forbus, 2010).

Multimodal object descriptions have been mostly studied in the context of multi-modal reference that typically involves pointing gestures, gaze, or iconic gestures. Existing computational models for multimodal REG have focussed on pointing and proposed different ways of combining or integrating verbal and deictic attributes: Kranstedt and Wachsmuth (2005) extend the classical incremental algorithm by Dale and Reiter (1995) to multi-modal attributes, such that the discriminatory power of the gesture determines the verbal content of the RE. Similarly, Van der Sluis and Krahmer (2007) assumes that deictic gestures are associated with a certain cost such that there is a certain competition between gesture and verbal content. Gatt and Paggio (2014) shows that the occurrence of pointing is tightly coupled with the RE's verbal realisation, based on data that records natural multimodal referring expressions. In this work, due to the lack of natural multimodal corpora, we leave out the task of learning temporal relations between verbal descriptions and sketches, but focus on investigating the effectiveness of combining verbal utterances with reduced sketches.

3 Task and Framework

Given a photograph of an object, we aim to generate multimodal descriptions composed of a verbal utterance and iconic information represented as sketch strokes (as shown in Figure 1), to enable a listener to identify a target object. As the *Draw-and-Tell* data does not reflect how human speakers would use sketch and verbal expression in combination, we implemented a straightforward baseline model that treats the two modalities as two independent channels: we split the multimodal generation task into two subtasks: *verbal description generation* and *stroke generation*. Formally, given a real-life photo $\mathcal{P}$, we aim to learn a model f that generates a description composed of an utterance $\mathcal{U}$ and sketch strokes $\mathcal{S}$:

$$f : \mathcal{P} \rightarrow (\mathcal{U} \times \mathcal{S}) \tag{1}$$

We opted for a simple, parallel architecture that takes the visual features of a photograph as input, and generates verbal descriptions and sketch strokes with separate models. While we adopted two mainstream natural language generation models for the verbal description generation (see Section 3.1), we went for a rule-based stroke generation approach with simplifying assumptions concerning the given data: Instead of generating the sketch in an end-to-end way, we cast it as a selection task where single strokes are extracted from the human hand-drawn sketches provided by the corpus, which we assume a computer vision module would generate with the given image in an end-to-end system.

Although it seems a simple edge detection model such as Canny (1987) would provide object edges to represent iconic information, note that, besides iconic information, hand-drawn sketches also reflect abstract, salient scene structures that people preserve when sketching from memory (Brady et al., 2008),
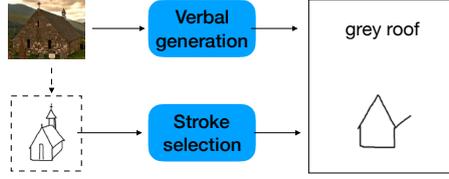
Figure 2: Framework of the multimodal description generation model.

Figure 3: Top 5 ranked sketch strokes of the house in the target image in Figure 2. The two longest strokes almost demonstrates the contour of the house, while the rest enrich the details with shorter strokes.

and are visually more distorted than extracted boundaries. Estimating human-like sketch behaviours is a challenging task by itself. Moreover, a simple stroke selection approach provides us with a relatively straightforward way of controlling the amount of details encoded in sketch strokes. Hence, we leave it as future work to generate object sketches with real life photos and form a complete system of generating multimodal descriptions of real life photos.

3.1 Verbal description generation

State-of-the-art systems for image captioning or REG on real-world images mostly rely on data-intensive deep learning models, e.g. (Vinyals et al., 2015; Yu et al., 2017). In contrast to large-scale data sets available for image captioning, the *Draw-and-Tell* data is comparatively small, but, at the same time, has a very large vocabulary with low-frequency words (see Section 4). Therefore, along with an RNN model, we also tested a retrieval-based model for generating object descriptions, which will be combined with iconic elements to form multimodal descriptions.

Recurrent neural network (RNN) generator We train a standard RNN for image captioning, as provided by Tanti et al. (2017). We use their inject architecture, which inserts the visual vector of the image at each time step of the RNN and predicts a distribution of the vocabulary. The hidden layer size is 256, and training was done for 3 epochs. The generator does not have an explicit representation of the distractors in the scenes. That is, it only considers visual features of target objects. We experimented with adding context features as in (Yu et al., 2017) or a discriminative loss function as in (Mao et al., 2016). However, this severely decreased the performance of the RNN, which is probably due to data sparsity.

Retrieval-based generation Alternatively, we implemented a simple consensus-based model of nearest neighbour retrieval, which can produce near state-of-the-art results in image description generation (Devlin et al., 2015). The generation algorithm works as follows: We preprocess long captions produced by humans into single phrases using commas, conjunctions and prepositions (e.g., *with*) as phrase marks. For a given test image, we retrieve its K nearest neighbours from the training set (K was tuned on the validation set). All phrases of k nearest neighbours are considered as candidate phrases and were ranked according to their consensus (Devlin et al., 2015), which is computed for each candidate phrase as the average word overlap **F-score** with all other candidate phrases. The top-ranked output phrase contains words that appear in many other expressions produced for the nearest neighbour images.

3.2 Stroke selection from hand-drawn sketches

As sketches in the Draw-and-Tell data are composed of single strokes, the stroke selection can be implemented in a straightforward way. Unfortunately, we do not have insights into how humans would sketch objects in an actual dialogue. Intuitively, drawing a full sketch as in Figure 1 would be too

time-consuming and inefficient in real-time interactions under timing constraints. We designed a simple rule-based stroke selection strategy based on the following criteria:

- **Simplicity**: the strokes must be simple, so that they can be easily drawn in a human-like manner in interactions. We observed that, to drawn the same length, a long stroke that can be drawn continually is less time consuming and looks more natural than a couple of short strokes.

- **Informativeness**: Each stroke should be informative so that a human listener can interpret the sketch by comparing the stroke with object parts in photos. For example, by comparing a stroke to the contour of a church, a listener should recognise that the stroke represents e.g. the roof of the church, rather than the walls. We observed that long strokes are often more visually salient and more informative than short strokes.

Considering the above observations, we selected the two longest strokes in each sketch to represent the corresponding object. Technically, we first parsed the SVG file of a sketch to a set of strokes and computed the length of each stroke with the Svgpath package.[1] The stroke length was calculated by recursive straight line approximations. After segmenting each stroke into at least 32 smaller segments, we took the sum of lengths of all segments as the stroke length. Then we ranked the strokes according to the lengths and select the two longest, as shown in Figure 3. On average, the two longest strokes in each sketch accounted for 40% of the total stroke length, with a standard deviation of 0.22.

4   Data

We used the *Draw-and-Tell* corpus (Han and Schlangen, 2017) to build and evaluate our generation models. The corpus contains 10,805 photographs which were selected from the ImageNet dataset (Russakovsky et al., 2015) and spanning over 125 categories. Each photo contains a single object, and was paired with a natural language description as well as several hand-drawn sketches (as shown in Figure 1).

The verbal descriptions were collected with an annotation task, with instructions similar to a reference task. Humans were asked to list all the attributes that can distinguish the target object from 6 other images in the same category, aiming to elicit fine-grained descriptions of visual attributes. Object attributes such as orientation, colour, shape, size, as well as any other attribute that might be helpful were suggested to be described (for more details, please refer to the original paper). Therefore, the descriptions often contain several short phrases of attribute descriptions (e.g., *facing leftwards, wet body*).

In addition, each photo is paired with around 5 different hand-drawn sketches derived from the Sketchy dataset (Sangkloy et al., 2016). The sketches were collected from non-professional workers. In other words, they represent sketching behaviours of average people. The hand-drawn sketches were saved as SVG files with high resolution timing information and stroke path information. This enables us to decompose sketches into single strokes.

Data statistics   On average, each description contains 2.79 phrases (separated by commas). The *Draw-and-Tell* corpus came with a train-test split setup, with 9734 photos in the training set and 1071 photos in the test set. The training set has a vocabulary size of 4758. Among all words in the training vocabulary, 3382 words (70.1%) appear fewer than 5 times. Compared to the training set, the test set has a smaller vocabulary which only contains 1601 words. Moreover, among the 1601 words, 224 words (14.0%) are not included in the training set vocabulary, making it a very challenging data set for learning to generate descriptions directly from visual input: a large vocabulary in the training set, many unknown words in the test set, and objects with similar visual attributes which can be difficult to describe with words.

5   Results

Using the train-test split setup in the *Draw and Tell* corpus, we trained two models for verbal description generation and implemented the stroke selection strategy as in Section 3.2. In order to generally estimate the quality of generated verbal descriptions, we performed automatic evaluation on the full test set. We

---

[1]`https://pypi.python.org/pypi/svg.path`

Models	F1-score	Precision	Recall	Av. length	Vocabulary size
Retrieval	0.24	0.355	0.205	4.78	135
RNN	0.176	0.204	0.167	7.96	114
Human	-	-	-	9.18	337

Table 1: Word overlap between generated and human descriptions for RNN and Retrieval system.

also conducted a task-based evaluation for unimodal and multimodal descriptions via crowdsourcing. For this, we randomly selected 100 <photo, sketch, description> pairs from the test set for evaluation.

5.1 Automatic NLG evaluation

First, we tested to what extent the generated verbal descriptions match human verbal descriptions, by computing the average word overlap between original and generated descriptions. As shown in Table 1, the retrieval-based model outperforms the RNN model by achieving higher **precision** and **recall** scores, although it generates much shorter descriptions on average. In other words, descriptions generated by the retrieval-based method are more precise and have a higher chance of mentioning an exact attribute of the target object, but might be too short and too ambiguous for discriminating the target from its distractor objects. The RNN is trained on full descriptions and produces longer descriptions, which are relatively less precise. This confirms the observation explained in Section 3.1 that the *Draw-and-Tell* corpus is challenging for data-intensive deep learning models. We also tested a retrieval-based method that generates longer descriptions, but found the F-score decreases rapidly when retrieving more than 1 phrase. Therefore, in the following, we focus on the RNN and retrieval-based system discussed above.

5.2 Task-based evaluation

We conducted a human evaluation with an object identification task. For each photo in the test set, we randomly selected 4 photos in the same category as distractor photos, forming a candidate set of 5 photos for each object identification task.

Experiment setup The experiments were conducted on the crowdsourcing platform Crowdflower[2]. As shown in Table 2, we ran 5 experiments with different combinations of NLG models and sketches. In each task, workers were asked to identify the target object from the range of candidate photos with a given unimodal or multimodal description. As this is a forced choice task, we additionally asked them to rate the confidence of their decision by clicking one of the four buttons: *random guess (0), uncertain (1), a bit uncertain (2), certain (3)*. To ensure the quality of the judgements, workers must complete a couple of test questions at first, which were derived from gold-standard descriptions in the corpus.

We presented generated descriptions and candidate photos in combination to workers. The candidate photos were shown in a row under each description. Workers were told that these descriptions were generated by a baby robot, who is learning to describe objects accurately and needs feedbacks about how accurate the descriptions are. We decided to contextualise the task in this way, to let workers know that the presented descriptions are not as accurate as those in standard annotation tasks (as most other tasks on Crowdflower). They were instructed to look at/read the descriptions, then look at the candidate photos, and select the ones that fits best with the descriptions by clicking the checkbox under the target photos.

We are aware of the fact that this is a rather simplified version of explanations as they are likely occur in dialogue, where the target object might not be physically present at the time of sketching (otherwise speakers might rather point to it). We leave it for future work to implement a more realistic version that temporally separates the presentation of description and real-world objects.

5.2.1 Human evaluation results

Table 2 shows the accuracy achieved by humans in the object identification task, along with average confidence scores. Overall, the sketch only setup achieves best performance; combining retrieval-based

---

[2]https://www.crowdflower.com

	-sketch	+sketch
-NLG	-	**0.53** / 2.14
+NLG-Retrieval	0.31 / 2.05	0.50 / **2.44**
+NLG-RNN	0.33 / 1.63	0.43 / 2.19
Chance level accuracy	0.20 / -	0.20 / -

Table 2: Human evaluation results. Object identification accuracy/confidence score for different combinations of NLG w/o sketch. For both metrics, a higher score indicates better performance.

descriptions with sketches marginally underperforms the sketch-only setup with a decreased accuracy by 0.03. In the **language-only** setup, the RNN model achieves a slightly higher accuracy score than the retrieval model. However, the uncertainty scores show that workers feel more confident about their decisions when reading descriptions generated by the retrieval model which are more human-like and grammatically correct, despite the fact that they are shorter than the RNN output. In the **sketch-only** setup, workers achieved the overall best performance with an accuracy score of 0.53, as well as a moderate confidence score of 2.14. Compared to the language-only experiment, this is a remarkable improvement in accuracy. Although the sketch strokes are often abstract, distorted and only contain limited details, they still effectively represent visual characteristics of the target objects.

RNN+Sketch > Sketch-only	Retrieval+Sketch > Sketch-only	Retrieval+Sketch > RNN+Sketch	Sketch only > Retrieval+Sketch
Retrieval: long white whiskers *RNN: gray and white in color facing left head facing right*	*Retrieval: white with orange beak* *RNN: white and black beak facing left*	*Retrieval: facing to the right side* *RNN: white and white in colour facing left facing left*	*Retrieval: coffee mug white colour* *RNN: white cup with white label on the table*
Retrieval: white body with red stripe *RNN: red and white in colour facing right*	*Retrieval: rabbit sits looking to the right on brown grass* *RNN: grey and white in colour facing left facing right*	*Retrieval: white shell* *RNN: a hermit crab is in the picture of the shell*	*Retrieval: the body is white* *RNN: white and white in colour has triangular shape*

Figure 4: Samples of generated descriptions, the head of the column indicates which system combination lead to a successful object description.

In the **multimodal** setup, both verbal generation models benefit from being combined with strokes. Interestingly, the improvement is much stronger for the retrieval-based model that generates shorter

descriptions. The multimodal retrieval model clearly outperforms the multimodal RNN system, even though the RNN is slightly better in the language-only condition. Overall, the multimodal retrieval model achieves the highest confidence scores. This results confirm previous findings on multi-modal embodied reference that it is effective for systems with imperfect perceptual capabilities (Fang et al., 2015). In the RNN-based system, however, language and iconic information seem to contradict each other to an extent that sketches are less effective. This also suggests that humans tend to put more weight on language descriptions.

Finally, it is note-worthy that multimodal descriptions slightly underperform the sketch-only descriptions in terms of accuracy. This further corroborates the observation that humans pay more attention to verbal descriptions, even if they are misleading.

For instance, we observed that utterances such as "*facing left*" and "*facing right*" are often confused by the NLG models, as they are probably not represented in current visual feature vectors and require ontological knowledge (i.e. where is the 'head of the object). However, this information about orientation is naturally represented in the sketches. These misleading verbal descriptions sometimes counterweigh the discriminative information encoded in sketch strokes. We conjecture that a multimodal description can even further improve the performance by modelling the interplay between the two modalities, and potentially restricting verbal descriptions to aspects that can be easily expressed symbolically (via words).

Qualitative examples for generated descriptions are shown in Figure 4. We made several observations here: the stroke selection strategy leads to iconic elements of very different quality. A human speaker might be unlikely to sketch a rabbit by only showing its hind leg, though this ultimately depends on the accompanying verbal expression. Other partial sketches clearly show the overall contour or shape of the object (e.g. the cat in column 1). Similarly, the verbal descriptions vary according to the properties they mention (colour, orientation, object parts) and according to their length. In contrast to strokes extracted from human sketches, verbal expressions are not always semantically adequate. The examples for multimodal descriptions outperforming unimodal ones seem to combine sketch and language in a complementary way where iconically signified properties relate to shape and verbally described properties mostly related to colour (column 2, 3).

6 Conclusion and Future Work

We take a first step towards generating multimodal object descriptions and propose to combine verbal expressions with iconic elements in the form of sketch strokes. Based on the *Draw-and-Tell* corpus, where verbal and sketched descriptions are available as parallel modalities, we implemented an RNN and a retrieval-based model to generate verbal descriptions for objects in real-life photographs, and selected sketch strokes from human sketches. The models were evaluated in a challenging object identification task, where fine-grained descriptions of visual attributes are essential for discriminating a target object from 5 distractors in the same category. The results show that descriptions combining sketch strokes with verbal descriptions not only achieve better performance than verbal descriptions, but also are perceived less confusing according to human ratings. Moreover, shorter descriptions from the retrieval-based model outperforms the RNN model when combined with sketches, indicating that short phrases together with sketches can be more effective than long but inaccurate verbal descriptions.

We believe that this work demonstrates the potential of using sketches for multimodal interaction and dialogue, even though we had to make some drastic simplifications in our setup and model. We found that even parallel unimodal data is useful for obtaining a baseline multimodal system. Yet, our results also clearly show that natural multimodal data is needed for modelling the interplay between iconic and verbal elements and get deeper insights into how these modalities convey meaning.

For future work, we plan to incorporate a computer vision module to automatically generate sketches from photos and work towards a real-time generation system presenting multimodal phrases in interactive setups, such as interactive referring games (Kazemzadeh et al., 2014). Moreover, as multimodal descriptions allow information to be expressed in two parallel modalities, they can be expected to allow for more efficient communication.

Acknowledgments

References

Aaron Adler and Randall Davis. 2007. Speech and sketching for multimodal design. In *ACM SIGGRAPH 2007 courses*, page 14. ACM.

Timothy F Brady, Talia Konkle, George A Alvarez, and Aude Oliva. 2008. Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, 105(38):14325–14329.

John Canny. 1987. A computational approach to edge detection. In *Readings in Computer Vision*, pages 184–203.

Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. 1994. Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pages 413–420. ACM.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.

Jan P De Ruiter, Adrian Bangerter, and Paula Dings. 2012. The interplay between gesture and speech in the production of referring expressions: Investigating the tradeoff hypothesis. *Topics in Cognitive Science*, 4(2):232–248.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proc. of CVPR*.

Jan de Wit, Thorsten Schodde, Bram Willemsen, Kirsten Bergmann, Mirjam de Haas, Stefan Kopp, Emiel Krahmer, and Paul Vogt. 2018. The Effect of a Robot's Gestures and Adaptive Tutoring on Children's Acquisition of Second Language Vocabularies.

Doug DeCarlo and Anthony Santella. 2002. Stylization and abstraction of photographs. In *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '02, pages 769–776, New York, NY, USA. ACM.

Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015. Language models for image captioning: The quirks and what works. In *Proceedings of ACL-IJCNLP 2015*, pages 100–105.

Rui Fang, Malcolm Doering, and Joyce Y Chai. 2015. Embodied collaborative referring expression generation in situated human-robot interaction. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 271–278. ACM.

Albert Gatt and Patrizia Paggio. 2014. Learning when to point: A data-driven approach. In *Proceedings of COLING 2014*, pages 2007–2017.

Bruce Gooch and Amy Gooch. 2001. *Non-photorealistic rendering*. AK Peters/CRC Press.

David Ha and Douglas Eck. 2018. A neural representation of sketch drawings. In *Sixth International Conference on Learning Representations*.

Ting Han and David Schlangen. 2017. Draw and Tell: Multimodal Descriptions Outperform Verbal- or Sketch-Only Descriptions in an Image Retrieval Task. In *Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP)*.

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating visual explanations. In *European Conference on Computer Vision*, pages 3–19. Springer.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2014. Referit game: Referring to objects in photographs of natural scenes. In *Conference on Empirical Methods in Natural Language Processing*.

Forbus Kenneth, Usher Jeffrey, Lovett Andrew, Lockwood Kate, and Wetzel Jon. 2011. Cogsketch: Sketch understanding for cognitive science research and for education. *Topics in Cognitive Science*, 3(4):648–666.

Stefan Kopp, Kirsten Bergmann, and Ipke Wachsmuth. 2008. Multimodal communication from multimodal thinkingtowards an integrated model of speech and gesture production. *International Journal of Semantic Computing*, 2(01):115–136.

Emiel Krahmer and Kees Van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Alfred Kranstedt and Ipke Wachsmuth. 2005. Incremental generation of multimodal deixis referring to objects. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*.

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

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

D McNeill. 1992. Hand and mind: What gestures reveal about thought. *What gestures reveal about*, pages 1–15.

Michael Oltmans and Randall Davis. 2001. Naturally conveyed explanations of device behavior. In *Proceedings of the 2001 workshop on Perceptive user interfaces*, pages 1–8. ACM.

Vaughan Prain and Bruce Waldrip. 2006. An exploratory study of teachers and students use of multimodal representations of concepts in primary science. *International Journal of Science Education*, 28(15):1843–1866.

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. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. 2016. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35 (4):119.

Rainer Stiefelhagen, C Fugen, R Gieselmann, Hartwig Holzapfel, Kai Nickel, and Alex Waibel. 2004. Natural human-robot interaction using speech, head pose and gestures. In *Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, volume 3, pages 2422–2427. IEEE.

Marc Tanti, Albert Gatt, and Kenneth Camilleri. 2017. What is the role of recurrent neural networks (rnns) in an image caption generator? In *Proceedings of INLG 2010*, pages 51–60.

Patrick Tresset and Frederic Fol Leymarie. 2013. Portrait drawing by paul the robot. *Computers & Graphics*, 37(5):348 – 363.

Barbara Tversky, Julie Heiser, Paul Lee, and Marie-Paule Daniel. 2009. Explanations in gesture, diagram, and word. *Spatial language and dialogue*, pages 119–131.

Barbara Tversky. 2014. Visualizing thought. In *Handbook of human centric visualization*, pages 3–40. Springer.

Ielka Van der Sluis and Emiel Krahmer. 2007. Generating multimodal references. *Discourse Processes*, 44(3):145–174.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Jon Wetzel and Ken Forbus. 2010. Design buddy: Providing feedback for sketched multi-modal causal explanations. In *Proceedings of the 24th International Workshop on Qualitative Reasoning*. Portland, Oregon.

Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. 2017. A joint speaker-listener-reinforcer model for referring expressions. Computer Vision and Pattern Recognition (CVPR). Vol. 2.

Sina Zarrieß and David Schlangen. 2016. Easy things first: Installments improve referring expression generation for objects in photographs. In *Proceedings of ACL 2016*, pages 610–620.