Draw and Tell: Multimodal Descriptions Outperform Verbal- or Sketch-Only Descriptions in an Image Retrieval Task

Anonymous IJCNLP submission

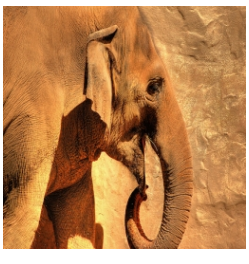
Abstract

While language conveys meaning largely symbolically, actual communication acts typically contain iconic elements as well: People gesture while they speak, or may even draw sketches while explaining something. Image retrieval *prima facie* seems like a task that could profit from combined symbolic and iconic reference, but it is typically set up to work either from language only, or via (iconic) sketches with no verbal contribution. Using a model of grounded language semantics on the one hand and a model of sketch-to-image mapping on the other, we show that even adding very reduced iconic information to a verbal image description improves recall. Verbal descriptions paired with fully detailed sketches still perform better than these sketches alone. We see these results as supporting the assumption that natural user interfaces should respond to multimodal input, where possible, rather than just language alone.

1 Introduction

In natural interactions, descriptions are typically multimodal: Someone explaining a route might point at visible landmarks while talking, or gesture them into the air, or may sketch a route on a piece of paper, if they have one handy (Emmorey et al., 2000; Tversky et al., 2009).

It is commonly assumed that these modes contribute to the joint meaning differently: the basis of the contribution of an utterance is conventional combination of conventional meanings (i.e., they contribute *symbolically*); pointing gestures contribute information *deictically* through a spatial connection to what they signify; other gestures, and sketches on paper, through similarity with what they represent (i.e., *iconically*) (Pierce, 1867; Kendon, 1980a; McNeill, 1992; Beattie and Shovelton, 1999).



Elephant, trunk coiled towards mouth, facing right

Figure 1: A photograph; a verbal description of its content; and a sketch

Work in computational semantics has mostly focussed on representing and composing symbolic information (Liang and Potts, 2015), with some recent attention to deictic information (Gatt and Paggio, 2013; Matuszek et al., 2014; Rautaray and Agrawal, 2015; Whitney et al., 2016; Han et al., 2015).

In this paper, we go beyond deictics, investigating iconic information in hand-drawn sketches which is often abstract and distorted. We address the question: to what degree iconic information—in our case here, coming from hand-drawn sketches of objects—can supplement symbolic information – verbal descriptions of objects.

We collected a corpus that pairs photographs with verbal descriptions and sketches (shown in Figure 1). The photographs were selected from ImageNet (Russakovsky et al., 2015), and paired with sketches in an existing corpus – the Sketchy Database (Sangkloy et al., 2016). We elicited verbal descriptions of objects in the photographs in a discriminative context (i.e., descriptions meant to single out the given object token in a set of related images). These descriptions provide us with

object attribute information which is either impossible to be sketched (e.g., colour in a monchrome sketch, material) and/or can easily be encoded in the symbolic mode (e.g., shape, orientation). In the former case, verbal descriptions complement iconicity in sketches, while in the latter case, verbal descriptions emphasise information potentially already in the sketch in a different communication modality.

We evaluated the joint contribution of iconic and symbolic information with an image retrieval task. To compose the meaning of multimodal descriptions, we use a recent model of grounded word meaning (the "Words as Classifiers" model (Kennington and Schlangen, 2015; Schlangen et al., 2016)) to evaluate how well a word fits with an image. We adopted the "triplet neural network" to evaluate the fitness between a given sketch and a photograph. And we evaluate the multimodal descriptions with a late fusion method, by combing the scores from the two models. By systematically reducing the level of details in the sketch, we investigate how much sketch detail can be recovered by verbal descriptions.

Our contributions are threefold: **a)** We introduce a corpus that pairs photographs with verbal descriptions and sketches, making it possible to investigate symbolic and iconic communications; **b)** We show that verbal descriptions and iconic information are supplementary. Enabling such multimodal input will lead to more informative expressions from humans, and could benefit other tasks such as reference resolution; **c)** We show that verbal descriptions and iconic information are also complementary. Verbal descriptions can make up loss of around 30% of sketch details. Enabling such iconic input could also reduce verbal effort.

The remainder of the paper is organised as follows: we first introduce the proposed dataset, and the data collection procedure; then we briefly describe the modelling of multimodal meaning, including comparing sketches with photographs and grounding verbal descriptions to photographs. Finally, we investigated the balance between verbal descriptions and sketches with controlled contributions from each modality. We discuss the results and the findings after describing each experiment.

2 The Data Set

We now first provide an overview of the Sketchy Database, then describe the experiment of pairing verbal descriptions with photographs, and the test set for evaluations in later sections.

2.1 Pairing Photographs with Sketches – The Sketchy Database

The Sketchy database contains 12500 unique photographs of real world objects which span 125 categories, and 75471 sketches from humans triggered by and paired with the photographs. A sketch-photograph pair is shown in Figure 1.

Photographs The 125 categories of photographs were selected based on the criteria that the objects in each category should be recognisable, specific and cover a large number of common objects. Each category should have recognisable sketch representations so that the sketches are not uninformative. Each of the photographs contains exactly one object and has bounding box annotations. Some categories of ImageNet photographs are combined into a single category to increase the visual diversity of photographs in the same category. The photographs were also graded with a subjective "sketchability" score. The 100 photographas in each category are with a target distribution of 40 very easy, 30 easy, 20 average and 10 hard.

Sketches The sketches were collected using Amazon Mechanical Turk (AMT)[1]. Workers were shown a random photograph from the database and instructed to sketch the named object with a similar pose to the object in the photograph on a canvas. They were instructed to only sketch the object and avoid shading regions. Workers clicked a button to view a photograph, which was shown for 2 seconds and hidden before participants started to sketch. Workers could view the photograph as many times as they want, however, after each viewing, the canvas was cleared. To make sure that the sketches are realistic and diverse, a visual noise mask was displayed after the photograph was hidden, so that low-level visual representations in visual working memory were masked. Therefore the sketches implicitly encode salient visual information of objects and are different from boundary annotations (Lin et al., 2014; Xiao et al., 2016).

Each sketch was stored as an SVG file which includes the start and end of stroke times and fine-grained timing along each stroke. (This en-

[1] https://www.mturk.com/mturk/welcome

Figure 2: Discriminative description of the left-most photograph provided by crowdworker: *facing right, trunk coiled toward mouth.*

ables us to reduce sketch details according to the timing information of strokes; see below in Section 4.2.) Five sketches, each from a different worker, were collected for each photograph. The photographs which were used to prompt sketches are paired with corresponding sketches and serve as gold standard photograph in the photograph retrieval task, which will be described below.

2.2 Pairing Photographs with Verbal Descriptions

To investigate the balance between iconic and semantic modes of object description, we paired the photographs with natural language descriptions.

Procedure We randomly selected 10,805 images from the Sketchy database (i.e., 85% of the whole database) and paired them with verbal descriptions. We used the "Crowdflower" service to collect the descriptions from English speakers. To elicit descriptions that convey salient attribute information of the named objects, we designed the job as a task to distinguish a named object from distractor objects in the same category (and not one of describing a single given image).

Workers were shown a photograph with an object and 6 other photos which are in the same category as the target photo, but visually different. They were asked to list a set of visual properties that can make the named subject photo distinguishable from the other 6 photos, separating each property with a comma. For example, Figure 2 shows an example of identifying an elephant from 6 distractor photographs of elephant. In this case, a worker described the shape of the elephant's nose and its orientation. (The actual interface presented to the workers showed the target image larger and on one side and the distractor objects in rows of 3 on the side; not shown here for reasons of space.)

We instructed workers to consider all and any attributes that help to distinguish the target object. Attribute types like colour, shape, material, orientation were suggested to workers, however, they were told to list any properties that they notice and can help another person to correctly select the right photograph from all the seven photographs.

In the following, we refer to these collected descriptions as "attributes" (att), to distinguish from the "category" (cat), which is known from the Sketchy database (and hence represent gold standard annotation without variation, as compared to the attributes).

To validate how informative the collected descriptions are, we randomly selected 100 of the descriptions and presented them to workers on Crowdflower as the 1-out-of-7 selection task. Workers correctly recognised 71% of the described photographs, which shows us that most of the verbal descriptions are informative, but some of them are ambiguous even for humans. (This can already be seen from the example in Figure 2: While the target photo might be the most salient one for which the description is true, it also matches the third one from the right.)

Data Statistics In total, we collected 10,805 object descriptions. 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.

Normalisation For grounding of terms to perceptual input (see below), we need a certain number of training instances (we set the cutoff at 10, see below). As the numbers reported in the previous paragraph show (and the fact that, as can be expected, the actual type distribution is Zipfian), this criterion would greatly reduce the number of useable descriptions. Moreover, the property specifications—each containing 4.6 words on average—have a compositional structure ("light blue", "trunk coiled toward mouth"), which our simple model of grounded semantic does not yet cover. We hence implemented a

rule-based normalisation step, which mapped various ways of expressing orientation (e.g., "facing to the left", "facing left", "looking to the left", "leftward looking") and colour ("of green colour", "green colored", "green") to a normalised representation (`facing_left`, `green`). After normalisation, we treat all properties as single tokens (i.e., we treat "trunk coiled toward mouth" as `trunk_coiled_toward_mouth`). In total, there are 18637 normalised properties. The ratio between types and tokens is 0.64. On average, each normalised property spans 3.41 words.

2.3 Testset

The Sketchy database comes with a suggested split into train and test sets. As we use the pre-trained networks, we follow this split.

The training set includes 9734 unique photographs, while the test set includes 1071 photographs (spanning all 125 categories). On average, each category includes 8.57 photographs in the test set; chance level accuracy of the sketch-image retrieval ($@K = 1$) task described below is 0.093%.

3 Modeling Multimodal Meaning

We compose the meaning of the multimodal description (sketch plus verbal description) out of the individual contributions; i.e., we perform what in the human/computer interaction community is called *late fusion* (Atrey et al., 2010).

3.1 Comparing Sketch and Photograph (Sangkloy et al., 2016)

We use the "Triplet Network" model introduced by Sangkloy et al. (2016) to compare sketches and photographs. The model maps sketches and photographs into a shared 1024-dimension embedding space.

More specifically, the triplet network includes a pair of deep convolutional networks (Szegedy et al., 2015) implemented in Caffe (Jia et al., 2014). One of the network accepts sketches as input (referred as sketch-network), while the other accepts photographs as input (referred as photo-network). The network uses a ranking loss function, with input tuples of the form (S, I+, I-) corresponding to a sketch, a matching image and a non-matching image. As a result, the network has a set of parameters for the sketch-network and a set of parameters for the photo-network. We used the two networks to map sketches and photographs in the test set into vectors. The distances between sketch and photo vectors indicates the semantic similarity between sketches and photographs. The smaller the distance, the better a sketch matches a photograph. We take the reciprocal of the distances as scores which show the appropriateness between sketches and photographs:

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

$\mathbf{S}$ indicates a feature vector of a sketch computed with the sketch network, and $\mathbf{P}$ indicates a feature vector of a photograph computed with the photograph network.

3.2 Grounding Verbal Descriptions to Photographs

We adopt the WAC ("words-as-classifiers") model (Kennington and Schlangen, 2015; Schlangen et al., 2016) to predict semantic appropriateness between words and referents in photographs. The model was originally introduced in a reference resolution task in dialogues, which is similar to our photograph retrieval task.

The WAC model pairs each word w (or here, each normalised attribute a) with a logistic regression classifier (trained with $\ell 1$ regularisation) which provides an "appropriateness score", given a vector of real-valued visual properties.

For example, to train a classifier for the word "elephant", we take all the photographs which contains elephants as positive training instances and randomly sampled the same amount of photographs which don't contain elephants as negative instances.

We trained classifiers for category and attribute words in the training dataset. With the trained word classifiers, first of all, we predict how well each word in a description fits with a candidate photograph, then we compose a fitness score with the scores of each word. For example, given a following description:

$$D : w_{a_1}, \cdots, w_{a_n}, w_c \quad (2)$$

where w_{a_i} indicates an attribute word, and w_c indicates a category word, we compute a score as following:

$$s_D(\mathbf{P}|D) = s_{cat}(\mathbf{P}|w_c) \times \sum_{i=1}^{n} s_{att}(\mathbf{P}|w_{a_i}) \quad (3)$$

where $s_{cat}(\mathbf{P}|w_c)$ indicates how well the category word fits with a photograph, and $s_{att}(\mathbf{P}|w_{a_i})$ indicates how well the attributes fit with a photograph. That is, we compose the attribute contribution additively, and it with the category representation multiplicatively (Schlangen et al., 2016).

The WAC classifiers were trained for the vocabulary of the training set. The minimum number of positive instances in the training set to train a WAC is 10. There are some words in the test set which were not included in the training set, in these cases, we set the response as 0 for each candidate image, thus the word doesn't contribute.

3.3 Fusion

Finally, for the full model, the scores were combined as following:

$$s_{sk+cat+att} = s_{sk}(\mathbf{P}|\mathbf{S}) \times s_D(\mathbf{P}|D) \quad (4)$$

4 Experiments

The photograph retrieval task We evaluated the contributions of sketches, category words, and attribute words with an image retrieving task. The goal is to retrieve a target photograph from the 1,071 photographs in the test set. We mapped the photographs into a 1,024 dimensional space for query. Each query can be made by a sketch image, or a verbal description, or together as described later in Section 4.3. Table 1 provides an overview of the models evaluated in following sections.

Metric Following the convention of image retrieval task evaluation, we measure performances of photograph retrieval models by recall @ *K*. For a given photo query, recall @ *K* is 1 if the corresponding photograph is among the top *K* retrieved results and 0 otherwise. We average over all test queries to produce average recalls. Here, we report the average recalls at $K = 1$ and $K = 10$.

4.1 Experiment 1: Mono-Modal Baselines

First of all, we inspect how well sketches and verbal descriptions can encode semantics by themselves with 3 mono-modal baseline models.

Baseline model 1: sk We use as baseline the performance of the original Sketchy model in the original setup (retrieval with sketch) as reported in (Sangkloy et al., 2016). This model reaches a performance of 0.36. (We re-ran the evaluation and reach numbers that are within 0.01 from the numbers reported in that original paper. This performance is plotted in Figure 4 as "sk", "100% of sketch".

Experiments	
Mono-modal baselines	sk
	cat
	att
Multimodal models	sk + cat
	sk + att
	sk + cat + att

Table 1: Image retrieving models with controlled input from each modality. (Retrieving with only sketches (sk), sketch + attribute words (sk + att), sketch + category words (sk+cat) and sketch + category words + attribute words (sk+cat+att)).

Baseline model 2: cat For the category word only evaluation, we only provide the category word as the description of each photograph. We applied the WAC classifier to compute a score for each candidate photograph, then take the one with highest score as the predicted photograph.

With category words only, we see an average recall @ 1 of 0.12 (@10: 0.90). As there are 8.57 photographs per category in the test set, ideally, given the category name, the chance level real @1 is 0.12, equal to the cat result. It shows that the category word classifiers performs well on distinguishing objects across categories, but not within a category.

Baseline model 3: att For the attribute words only evaluation, we apply WAC to each attribute word in a description. Each WAC classifier computes a score for all candidate photographs. We add up all the scores and take the one with the highest score as the most likely described photograph.

With attribute words only, @ 1 is 0.04 (@10: 0.23). These are shown in Figure 4, as "cat" and "att", respectively. Category words perform well above attribute words. The performance is well above the chance level 0.093%, however, it's significantly lower than the cat model. It suggests that attributes are neither category nor object specific. Thus it only prunes the retrieving results within a small domain (distinguish objects that are not facing right from those "facing right" objects.)

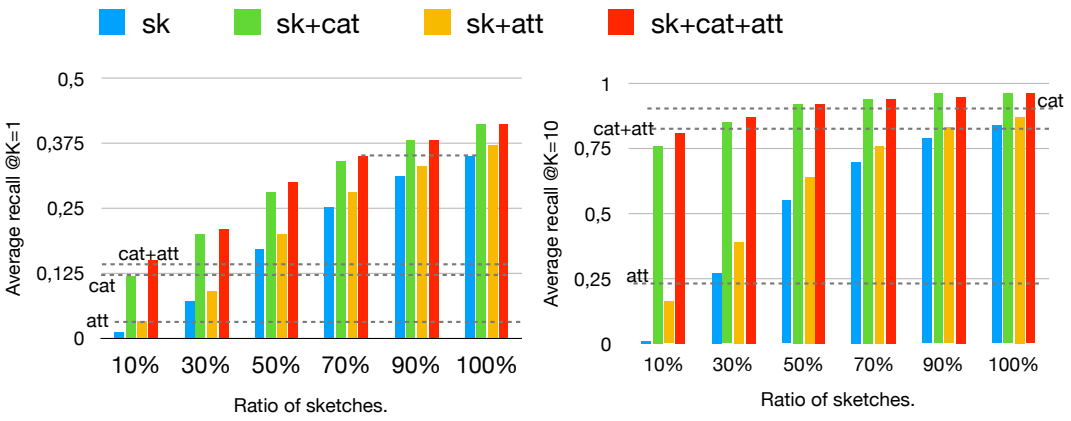
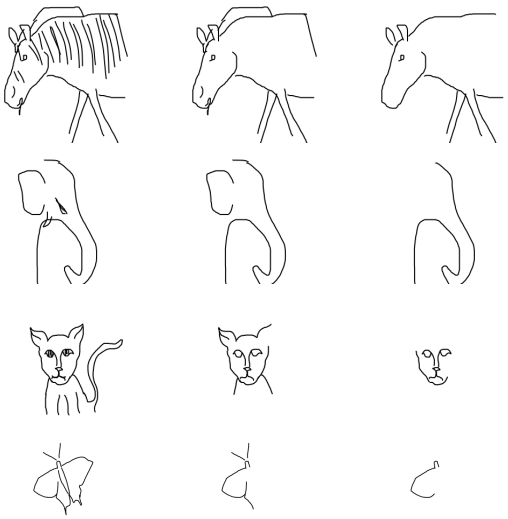


Figure 4: Average recall at K=1 and K=10..



(a) 100% (b) 50% (c) 30%

Figure 3: Full sketches and reduced sketches at different ratios.

4.2 Reducing Sketch Detail

While sketches in the Sketchy database were drawn from memory, they were instructed to sketch the specific given object instance. Thus the sketches encode rich details of object structures. To investigate how much detail sketches need to convey informative iconicity, we reduced the amount of strokes in each sketch and evaluated the performance of photo retrieval task on these reduced sketches.

As aforementioned, the sketches were stored as SVG files which includes high-resolution timing information. With the provided start and end times of strokes, we reduced sketches by leaving only the first 10% (30%, ...) of the strokes. Figure 3 shows some examples of this process. As these examples show, in many cases the early strokes already capture some salient information, with later strokes filling in more detail. We render the reduced SVG files to images and present those to the Triplet Network to extract features.

The "sk" bars in Figure 4 shows the evaluation results when retrieving photographs using reduced sketches. As can be seen, performance degrades near linearly with the reduction of sketch detail.

4.3 Experiment 2: Multimodal Retrieving

To investigate the balance between symbolic and iconic modes in the task, we designed experiments to retrieve photographs with both verbal descriptions and sketches. Detailed numbers of the experiments are reported in Table 2.

sk+cat First of all, we evaluated the performance when sketch and category words are jointly used. As shown in Figure 4, category words efficiently reduces the range of candidate pho-

tographs. The average recall @10 increased 0.1 after adding category words (@10: 0.71), when sketches are reduced to only 10%. With full sketches, the average recall @1 increased 0.06 (@10: 0.08). Hence, although sketches are object specific, category specific information from category words supplements sketches in the symbolic mode.

sk+att We also evaluated the performance when sketch and attribute words are jointly used. As shown in Figure 4, attribute words together with sketch slightly improve the retrieval performance. Although the improvement is less significant as category words, it's rather consistent while the sketch details are reduced. This indicates that attributes contribute more at lower detail settings and supplement the sketches. For example, both sketch and attribute words can signify the orientation "facing right". However, sketches can not encode colour attributes like "red". Thus attribute words supplement sketches. However, since colour attributes are neither object nor category specific, the improvement is limited.

sk+cat+att Finally, we evaluated the performance when sketch, category words and attributes are all applied to the photograph retrieval task. By combing all the information, the model reaches its best performance.

For evaluation among top 1 image, the contribution of category and attribute words seems somewhat less pronounced, with less than 80% sketch can achieve the same performance as a full sketch. By providing category information, performance at the reduction 10% goes up from 0.01 to 0.12. For evaluation among top 10 images (recall @ 10), 10% sketch accompanied with category and attribute words on average achieves equal performance to a much more detailed sketch (90%) without accompanying verbal information.

Figure 5 shows some examples of how symbolic and iconic semantics supplement each other. In Figure 5h, the sketch of an butterfly was reduced to only 30%. As shown, the remained strokes only shows the contour of one of the wings, which is not informative enough to be distinguished as a butterfly. As a result, the target photograph only got a rank of 32. In comparison, Figure 5g provides the details of another wing and the head, which enables our model to rank the target photograph as the most likely candidate ($rank$ = 1). We added verbal descriptions to the reduced sketch and retrieve with both sketch and words, the words provide category and attribute information, which leads to a good performance as when retrieving with a full sketch.

5 Related Work

Multimodal communications, especially language and gesture related natural communications, have been widely studied in recent years (Kendon, 1980b; Alibali, 2005; Kendon, 1980a). (McNeill, 1992) studied how gestures reveal the thoughts in our mind, and proposed that speakers reveal in their gestures what they regard as relevant and salient in the context.

While we are not aware of directly comparable computational work on multimodal ensembles consisting of iconic gestures/sketches and verbal descriptions, there is related work on multimodal descriptions in general. (Lücking et al., 2010) presented a corpus built of speech and gesture in a route description task. (Sowa and Wachsmuth, 2003, 2009) investigated coverable iconic gestures for object descriptions. The empirical study shows that gestures convey geometric attributes by abstraction from the complete object shape.

(Bergmann and Kopp, 2006; Allwood et al., 2016) investigated how language and gestures' semantics related to each other. While their work sheds light on how iconic gestures encode semantic information, they didn't systematically evaluate the interplay between iconic gestures and verbal content.

(Johnston et al., 2002; Stiefelhagen et al., 2004; McGuire et al., 2002; Matuszek et al., 2014) have shown that incorporating gestures in language related human-machine interaction systems can help to improve the performance. However, these works mainly focus on evaluate the improvement of system performance, rather than the interplay of semantics between modalities. In this work, we investigate how much details iconic gestures require to convey informative information and to what degree language can recover the reduced iconicity in gestures.

6 Conclusions

We presented a corpus that pairs photographs with verbal descriptions, and hand-drawn sketches from an existing database. With the corpus, we investigated the balance between symbolic and

Detail	10%		30%		50%		70%		90%		100%	
	@1	@10	@1	@10	@1	@10	@1	@10	@1	@10	@1	@10
sk	0.01	0.06	0.07	0.27	0.17	0.55	0.25	0.70	0.31	0.79	0.35	0.84
att	0.03	0.23	0.03	0.23	0.03	0.23	0.03	0.23	0.03	0.23	0.03	0.23
cat	0.116	0.90	0.116	0.90	0.116	0.90	0.116	0.90	0.116	0.90	0.116	0.90
cat+att	0.14	0.83	0.14	0.83	0.14	0.83	0.14	0.83	0.14	0.83	0.14	0.83
sk+att	0.03	0.16	0.09	0.39	0.20	0.64	0.28	0.76	0.33	0.83	0.37	0.87
sk + cat	0.12	0.76	0.20	0.85	0.28	0.92	0.34	0.94	0.38	0.96	0.41	0.96
sk + cat + att	0.15	0.81	0.21	0.87	0.30	0.92	0.35	0.94	0.38	0.95	0.41	0.96

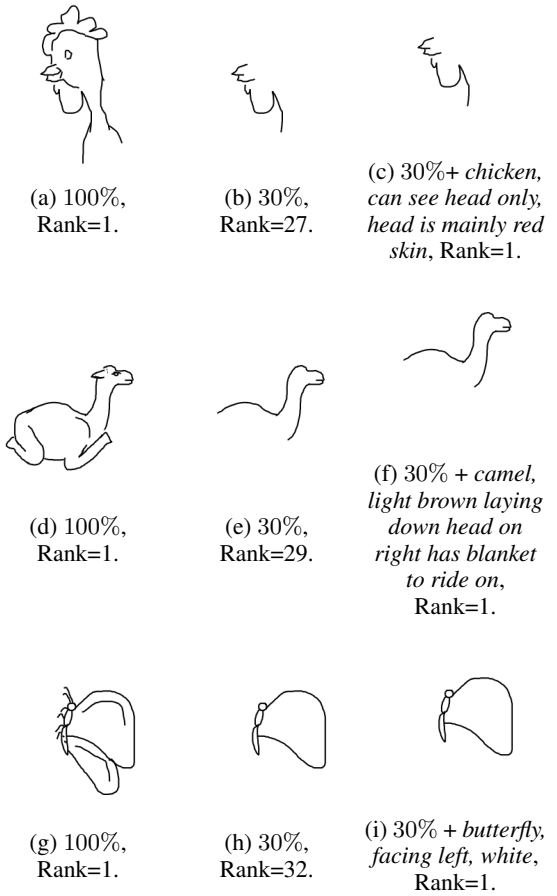
Table 2: Average recall at K=1 and10, with different detailed sketches. (Red numbers are the maximum recalls in each column.)

iconic modes in object descriptions with controlled contributions in each modality (by reducing sketch details, or the type of information provided by the symbolic mode). Evaluating in a photo retrieval task (as a stand-in for the task of recognising the referent of a multi-modal description), we showed that iconic information represented in a continuous vector space can be combined with verbal information through late fusion.

Moreover, we showed that the iconic mode and the symbolic mode indeed carry complementary/supplementary information. While category words and attribute words distinguish images across and within categories respectively, sketches add more information through visual similarities. The more details in the sketches, the more they contribute. We believe that enabling such multimodal communications will lead to more informative descriptions, and reduce cognitive load from each modality. Finally, we will make the corpus publicly available in the future.

7 Future work

While we used hand-drawn sketches in this work, our ultimate interest is in interpreting iconic gestures. Sketches are similar to gestures in the sense that both signify iconicity in a visual way. 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 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



(a) 100%, Rank=1.
(b) 30%, Rank=27.
(c) 30%+ *chicken, can see head only, head is mainly red skin*, Rank=1.
(d) 100%, Rank=1.
(e) 30%, Rank=29.
(f) 30% + *camel, light brown laying down head on right has blanket to ride on*, Rank=1.
(g) 100%, Rank=1.
(h) 30%, Rank=32.
(i) 30% + *butterfly, facing left, white*, Rank=1.

Figure 5: Photo retrieving with 100% sketch, 30% sketch, and 30% sketch + verbal description.
[das: Add ranks for verbal-only.]

are more close 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 apply related methods to iconic gesture, where efforts are under way to provide comparable data using motion capture data as input. (A considerable effort, as existing gesture data sets typically only provide small sets of very abstract, acted gestures, e.g. (Liu and Shao, 2013). We believe that we are at a good starting point for that with the work reported here.

References

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

Jens Allwood, Elisabeth Ahlse, et al. 2016. Meaning potentials in words and gesture. In *Proceedings from the 3rd European Symposium on Multimodal Communication, Dublin, September 17-18, 2015*. Linköping University Electronic Press, 105, pages 1–6.

Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems* 16(6):345–379.

Geoffrey Beattie and Heather Shovelton. 1999. 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.

Kirsten Bergmann and Stefan Kopp. 2006. Verbal or visual? how information is distributed across speech and gesture in spatial dialog. In *Proceedings of the 10th Workshop on the Semantics and Pragmatics of DIalogue*. pages 90–97.

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

Albert Gatt and Patrizia Paggio. 2013. What and where: An empirical investigation of pointing gestures and descriptions in multimodal referring actions. In *ENLG*. pages 82–91.

Ting Han, Casey Kennington, and David Schlangen. 2015. 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)*.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, pages 675–678.

Michael Johnston, Srinivas Bangalore, Gunaranjan Vasireddy, Amanda Stent, Patrick Ehlen, Marilyn Walker, Steve Whittaker, and Preetam Maloor. 2002. MATCH: An architecture for multimodal dialogue systems. *Computational Linguistics* pages 376–383.

Adam Kendon. 1980a. Gesticulation and speech: Two aspects of the process of utterance. *The relationship of verbal and nonverbal communication* 25(1980):207–227.

Adam Kendon. 1980b. Gesticulation and speech: two aspects of the process of utterance. *The Relationship of Verbal and Nonverbal Communication* 25:207–227.

Casey Kennington and David Schlangen. 2015. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. In *Proceedings of the Conference for the Association for Computational Linguistics (ACL)*.

Percy Liang and Christopher Potts. 2015. Bringing machine learning and compositional semantics together. *Annual Review of Linguistics* pages 1–27. https://doi.org/10.1146/annurev-linguist-030514-125312.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*. Springer, pages 740–755.

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

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*.

Cynthia Matuszek, Liefeng Bo, Luke Zettlemoyer, and Dieter Fox. 2014. Learning from Unscripted Deictic Gesture and Language for Human-Robot Interactions. In *AAAI 2014*.

P. McGuire, J. Fritsch, J.J. Steil, F. Rothling, G.a. Fink, S. Wachsmuth, G. Sagerer, and H. Ritter. 2002. Multi-modal human-machine communication for instructing robot grasping tasks. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*. volume 2, page 7.

D McNeill. 1992. Hand and Mind: What Gestures Reveal About Thought. *What gestures reveal about* pages 1–15.

Charles Sanders Pierce. 1867. 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.

Siddharth S. Rautaray and Anupam Agrawal. 2015. Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review* 43:1–54. https://doi.org/10.1007/s10462-012-9356-9.

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.

David Schlangen, Sina Zarrieß, and Casey Kennington. 2016. Resolving references to objects in photographs using the words-as-classifiers model. In *Proceedings of ACL 2016*. Berlin, Germany.

Timo Sowa and Ipke Wachsmuth. 2003. Coverbal iconic gestures for object descriptions in virtual environments: An empirical study. *Gestures. Meaning and Use* pages 365–376.

Timo Sowa and Ipke Wachsmuth. 2009. A computational model for the representation and processing of shape in coverbal iconic gestures. In *Spatial Language and Dialogue*.

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

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pages 1–9.

Barbara Tversky, Julie Heiser, Paul Lee, and MariePaule Daniel. 2009. Explanations in Gesture, Diagram, and Word. In Kenny R. Coventry, Thora Tenbrink, and John Bateman, editors, *Spatial Language and Dialogue*, Oxford University Press, pages 119–131.

David Whitney, Miles Eldon, John Oberlin, and Stefanie Tellex. 2016. Interpreting multimodal referring expressions in real time. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*. IEEE, pages 3331–3338.

Jianxiong Xiao, Krista A Ehinger, James Hays, Antonio Torralba, and Aude Oliva. 2016. Sun database: Exploring a large collection of scene categories. *International Journal of Computer Vision* 119(1):3–22.