SUN-Spot: An RGB-D Dataset With Spatial Referring Expressions

Cecilia Mauceri, Martha Palmer and Christoffer Heckman
University of Colorado Boulder
Department of Computer Science, 430 UCB
Boulder, Colorado 80309-0430
<first>.<last>@colorado.edu

Abstract

We introduce a new dataset, SUN-Spot, for localizing objects using spatial referring expressions (REs). SUN-Spot is the only RE dataset which uses RGB-D images. It also contains a greater average number of spatial prepositions and more cluttered scenes than previous RE datasets. Using a simple baseline, we show that including a depth channel in RE models can improve performance on both generation and comprehension.

1. Introduction

Spatial information can clarify ambiguous instructions and identify unknown objects. Humans prefer to use spatial information to differentiate objects even when they could choose other object descriptions such as color, shape, or size [22]. Therefore to develop more effective human-computer interaction, we need models of grounded spatial language. In this work, we focus on phrases which uniquely identify objects using spatial information, or spatial referring expressions (REs).

Spatial REs are challenging to model because they require understanding additional context. Appearance-based descriptions like color, shape, or object class, require detecting the attributes of the target object alone. In contrast, spatial descriptions require understanding the relationship between the landmark object and the target object. Additionally, spatial REs are often perspective-dependent.

To address these challenges, we introduce a new dataset, SUN-Spot, which combines RGB-D images with spatial REs. Depth is an increasingly ubiquitous sensing modality. Robots are typically equipped with depth sensors to support grasping, manipulation, and navigation. Mobile phones and personal computers are using depth sensing for facial recognition and augmented reality. Depth is also an important dimension for spatial language with "behind" and "in front" being among the top prepositions occurring in SUN-Spot. We hypothesize that including depth in spatial RE models will improve performance.

SUN-Spot contains 1948 images and 7987 REs, with an average of 2.6 spatial prepositions per expression. An example from our dataset is shown in Figure 1. Compared to existing REs datasets, this dataset has longer descriptions, more spatial prepositions, and is the only dataset including a depth channel. The full dataset is available at arpg.colorado.edu/sunspot.

2. Related Work

RE datasets with synthetic images have been used in NLP for the past decade to study the generation of REs [12]. More recently, interest in expanding the scope of Visual Question Answering (VQA) has produced several large scale data sets, both synthetic, such as CLEVR-Ref+ [15], and realistic, such as ReferIt [8] and Google RefExp (RefExp) [18]. Other closely related data sets include visual dialog systems [3, 4], where the goal is to generate a series of REs which zero in on one target object, and navigation data sets [1] which use REs to direct a robot to a goal.

SUN-Spot most resembles the RefExp and ReferIt datasets. It differs in three important ways: (1) the focus on spatial relationships between objects, (2) the composition of the images, and (3) the use of RGB-D images. SUN-Spot contains the highest mean location prepositions per RE (See Table 1). Furthermore, SUN-Spot images are keyframes from a video stream and are therefore more closely resemble the visual input of a mobile robot. Characteristics of the keyframes include bad lighting conditions, non-level camera frames, and a large amount of clutter. In contrast, the photos used by RefExp and ReferIt are gathered from photo collections on the web. RefExp and ReferIt image are usually well lit and have a small number of highly salient objects. However, the most important difference is that the SUN-Spot RGB-D images include a depth channel, while RefExp and ReferIt are RGB only. Depth has been shown to improve accuracy for scene understanding [16] and manipulation [2]. To our knowledge, our dataset is the only dataset which combines RE annotations for RGB-D images.

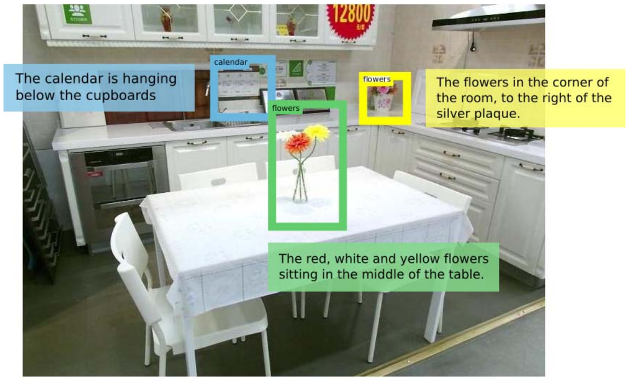
The calendar is hanging below the cupboards
calendar
flowers
flowers
The flowers in the corner of the room, to the right of the silver plaque.
The red, white and yellow flowers sitting in the middle of the table.

Figure 1: An example RGB image from the SUN-Spot dataset including three object bounding boxes with referring expressions superimposed

3. Data Set

The SUN-Spot dataset extends the scene understanding dataset SUN RGB-D [21]. SUN RGB-D contains 2D object segmentation and 3D object bounding boxes with orientation for over 10,000 RGB-D images of indoor scenes. A subset of 1449 images, originally the NYUv2 dataset [19], has been previously annotated with captions [11] and visual questions [17].

We annotated 1948 images with REs. The images were selected with a focus on images containing two or more objects from the same object class, similar to the methodology of RefExp [18]. The need to discriminate between objects of the same class within the same scene forces the annotators to provide more detailed descriptions. Unlike RefExp, we also include images with only one instance of the object to achieve a balanced distribution of the object classes occurring in the SUN RGB-D dataset. Our image selection process first computed the number of occurrences of each object class in each image. For each multiply-occurring object class, we then selected a random sample of 50 images containing at least 2 objects of that class. Some classes like "oven" never appear more than once in the same image. For these classes, we take a random sample of 10 images depicting these classes to avoid excluding object classes that appear in the SUN RGB-D dataset. For example, Figure 1 shows two labeled objects from the class "flowers" and one labeled object from the uncommon class "calendar."

Table 1 summarizes the size and complexity of the resulting dataset. It has long descriptions compared to other RE datasets and a larger average number of spatial prepositions per annotation. The most similar dataset in terms of expression length and frequency of location prepositions is RefExp. Figure 2 shows that eight of the ten most frequent location prepositions are shared by SUN-Spot and RefExp.

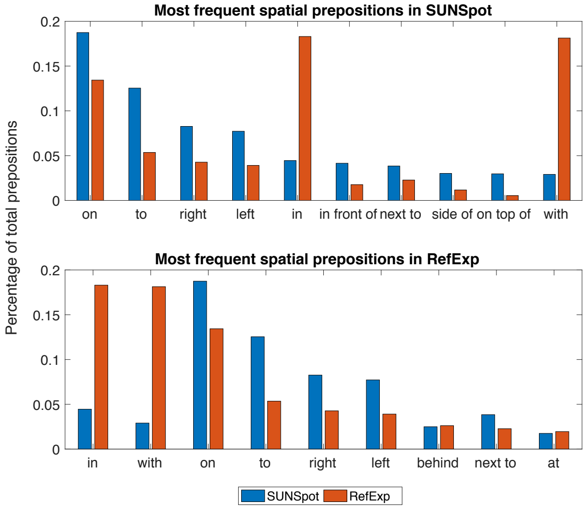

Figure 2: Relative frequency of the 10 most frequent location prepositions in RefExp and SUN-Spot

To calculate the frequency of location prepositions, we use a vocabulary of common location prepositions and count token occurrences in the dataset. Since these prepositions may have multiple senses, this calculates an upper-bound on their usage as location prepositions. Several prepositions that unambiguously refer to location, such as 'on,' 'left,' 'right,' 'front,' and 'top,' occur in greater proportion in SUN-Spot.

4. Experiments

In order to compare the challenges of learning on SUN-Spot to other RE datasets we run generation and comprehension experiments using the same baseline RE generation network as Mao *et al.* [18]. Furthermore, to investigate the value of SUN-Spot's depth channel, we modify the baseline slightly to accept depth input and compare to RefExp with synthetically generated depth.

4.1. Models

Mao *et al.*'s baseline network consists of two main components, an image network to encode the image features, followed by an LSTM to generate text. Across all experiments, the LSTM architecture remains the same, and we substitute two different image models described below. Furthermore, we modified Mao *et al.*'s training procedure and architecture in the following ways to improve training time and performance. First, we used L2-regularization on the LSTM weights. We disregarded dropout as we saw little to no improvement at cost of training time. We also used the Adam Optimizer [10] to train, where Mao *et al.* use vanilla stochastic gradient descent. Our implementation is available as a Github repository: https://github.com/crmauceri/ReferringExpressions.

Dataset	Images	REs	Vocab	Classes	Length	LocPrep
SUN-Spot	1,948	7,990	2,690	578	14.04	2.60
ReferIt [8]	19,997	130,364	9,320	276	3.51	0.76
RefCOCO [24]	19,994	142,209	10,341	80	3.50	0.87
Google RefExp [18]	25,799	95,010	2,890	80	8.41	1.23

Table 1: A comparison of RE datasets in terms of the number of images (Images), referring expressions (REs), average location prepositions per RE (LocPrep), average words per RE (Length), and number of unique object classes (Classes).

RGB Models For direct comparison between the SUN-Spot and RefExp, we omit SUN-Spot's depth channel so that the models can accept both SUN-Spot and RefExp examples as input. We use a pretrained VGG-16 network [20] to produce image features.

We trained two RGB RE networks. The first model, *Baseline*, was trained for 60 epochs on the RefExp dataset. For the second model, *Baseline+fine*, we fine-tuned the Baseline model with the SUN-Spot training set for a further 30 epochs.

RGB-D Models To test the potential gains from adding depth based features, we train a custom VGG-16 network with a 4th channel added to the first convolutional layer of a conventional VGG-16 network. Because no other RE dataset contains RGB-D images, we used synthetic depth. Using MegaDepth [13], we generated a synthetic depth channel for the COCO dataset [14], the source of images for RefExp. We train the VGG-16 network for 65 epochs on the portion of the COCO 2014 training set disjoint from the RefExp dataset. We use multi-label binary cross entropy loss to predict all the object labels in each image.

We train two RE models with depth. The first model, *VGG+D*, is trained for 30 epochs on the RefExp training set with the added depth channel. For the second model, *VGG+D+fine*, we fine-tune the VGG+D model on SUN-Spot for a further 5 epochs. For a direct comparison, we also train an RGB VGG-16 in the same way as the depth networks, *VGG* and *VGG+fine*.

We also experimented with HHA depth preprocessing [5] which is the standard approach for incorporating depth into image networks. However, we observed that HHA depth preprocessing magnified errors in surface normal prediction in synthetic depth images. Additionally, many of the COCO images do not have a ground-plane, which is required to calculate HHA. Therefore we did not find HHA suitable for synthetic depth.

4.2. Referring Expression Generation

We evaluate our generated expressions with automated metrics, BLEU, ROUGE-L, and CIDEr [9]. Traditionally used for measuring the quality of machine translation and image captioning, they can also be used for comparing the similarity of two REs. Table 2 shows a summary of the similarity metrics. The first two rows show the datasets' internal similarity across REs describing the same object. To calculate this value, we held out one expression from each set of expressions describing the same object. These scores can be considered upper bounds on what generated expressions can achieve as they represent the natural variance between human annotators. They also show that SUN-Spot has more internal similarity than RefExp by all three metrics. The difference in score between datasets confirms that RefExp and SUN-Spot do have significant biases that stymie transfer learning from one to the other for generating expressions. These biases could stem from different vocabulary or from different sentence structure. It is nevertheless impressive that the Baseline+Fine shows such improved performance on SUN-Spot despite the relatively small size of that dataset.

Model	Dataset	B1	R-L	C	P@1
-	RefExp	0.33	0.31	0.82	
-	SUN-Spot	0.56	0.51	1.33	
Baseline	RefExp	0.30	0.31	0.31	0.50
Baseline	SUN-Spot	0.27	0.19	0.07	0.20
Baseline+fine	RefExp	0.18	0.21	0.08	0.50
Baseline+fine	SUN-Spot	0.45	0.44	0.15	0.33

Table 2: Quantitative results for generation and comprehension on RGB models. Columns are BLEU1 (B1), ROUGE-L(R-L), CIDEr(C), and Precision at 1(P@1). The first two rows compare ground truth REs to establish an upper-bound. The other rows evaluate generated sentences.

4.3. Referring Expression Comprehension

The models generate REs, but we can also use them to measure the comprehension of REs by ranking the likelihood of generating the input expression. Generation likelihood ranking was introduced simultaneously by Mao *et al.* [18] and Hu *et al.* [6] and has been widely used since to use generative networks for comprehension [1, 24]. To compute generation likelihood, for each target RE, S, we select the bounding box, R^*, which maximizes the probability of generating the target expression for the given image I. This can be expressed as

$$R^* = \arg\max_{R \in C} p(R|S, I) \tag{1}$$

Model	Dataset	B1	R-L	C	P@1
VGG	RefExp	0.17	0.19	0.15	0.25
VGG+D	RefExp	0.18	0.20	0.21	0.25
VGG+fine	SUN-Spot	0.30	0.35	0.11	0.13
VGG+D+fine	SUN-Spot	0.34	0.34	0.14	0.17

Table 3: Results comparing RGB image features (VGG and VGG+fine) to RGB-D image features (VGG+D and VGG+D+fine). Metrics are the same as used in Table 2.

where C is the set of all bounding boxes in image I.

In a fully automated scenario, the bounding boxes would be generated by a bounding box proposal system. To estimate an upper-bound on the performance, we use the ground truth bounding boxes. As a metric, we use comprehension precision at 1(P@1), which measures whether the correct bounding box had the highest generation likelihood for a given expression. We compare 8 ground truth bounding boxes per image in the RefExp dataset and 10 bounding boxes per image in the SUN-Spot dataset. Randomly selecting a bounding box would yield 12% precision@1 for the RefExp dataset and 10% for the SUN-Spot dataset.

We report the precision for the comprehension task in Table 2. The precision for the RefExp dataset does not drop after fine-tuning. This suggests that the SUN-Spot fine-tuning leads to better generalization for the comprehension task.

4.4. Effects of Depth

Table 3 compares RGB and RGB-D models. The results for VGG and VGG+D are similar. The addition of a synthetic depth channel has a limited effect on performance. However, between the fine-tuned models, we see a significant improvement in VGG+D+fine, trained with real depth measurements. Real depth measurements, from SUN-Spot, improve both generation accuracy and comprehension precision. This is an interesting result as it underscores the value of RGB-D datasets in building multi-model RE models.

5. Conclusion

SUN-Spot is a new dataset focused on spatial expressions describing objects in cluttered interior scenes. It contains more objects per image, longer descriptions, and more location prepositions per description than competing RE datasets. It is the only RE dataset with RGB-D images. Using depth in multi-modal RE models improves both generation and comprehension.

Acknowledgments Research sponsored by DARPA Award: FA8750-18-2-0016. Thank you to Tim O'Gorman for providing a list of spatial prepositions.

References

[1] A. Balajee Vasudevan, D. Dai, and L. Van Gool. Object referring in visual scene with spoken language. In *WACV*, 2018. 1, 3

[2] J. Bohg et al. Data-driven grasp synthesisa survey. *IEEE Transactions on Robotics*, 30(2):289–309, April 2014. 1

[3] A. Das et al. Visual Dialog. In *CVPR*, 2017. 1

[4] H. De Vries et al. GuessWhat?! Visual object discovery through multi-modal dialogue. In *CVPR*, 2017. 1

[5] S. Gupta et al. Learning rich features from RGB-D images for object detection and segmentation. In *ECCV*, 2014. 3

[6] R. Hu et al. Natural Language Object Retrieval. In *CVPR*, 2016. 3

[7] A. Janoch et al. A Category-Level 3D Object Dataset: Putting the Kinect to Work. In *Consumer Depth Cameras for Computer Vision: Research Topics and Applications*, pages 141–165. Springer, London, 2013.

[8] S. Kazemzadeh et al. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *EMNLP*, 2014. 1, 3

[9] M. Kilickaya et al. Re-evaluating automatic metrics for image captioning. In *EACL*, 2017. 3

[10] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv*, 2014. 2

[11] C. Kong et al. What are you talking about? Text-to-Image Coreference. In *CVPR*, 2014. 2

[12] E. Krahmer and K. Van Deemter. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218, 2012. 1

[13] Z. Li and N. Snavely. MegaDepth: Learning single-view depth prediction from internet photos. In *CVPR*, 2018. 3

[14] T.-Y. Lin et al. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 3

[15] R. Liu et al. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *CVPR*, 2019. 1

[16] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1

[17] M. Malinowski, M. Rohrbach, and M. Fritz. Ask Your Neurons: A Deep Learning Approach to Visual Question Answering. In *International Journal of Computer Vision*, volume 125, pages 110–135, 2017. 2

[18] J. Mao et al. Generation and Comprehension of Unambiguous Object Descriptions. In *CVPR*, 2016. 1, 2, 3

[19] N. Silberman et al. Indoor segmentation and support inference from RGB-D images. In *ECCV*, 2012. 2

[20] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv*, 2014. 3

[21] S. Song et al. SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite. In *CVPR*, 2015. 2

[22] H. Viethen. *The Generation of Natural Descriptions: Corpus-based Investigations of Referring Expressions in Visual Domains*. Australasian Digital Theses Program. Macquarie University, 2011. 1

[23] J. Xiao, A. Owens, and A. Torralba. SUN3D: A Database of Big Spaces Reconstructed using SfM and Object Labels. In *CVPR*, 2013.

[24] L. Yu et al. Modeling context in referring expressions. In *ECCV*, 2016. 3