SpatialVOC2K: A Multilingual Dataset of Images with Annotations and Features for Spatial Relations between Objects

Anja Belz
Computing, Engineering and Mathematics
University of Brighton
Lewes Road, Brighton BN2 4GJ, UK
a.s.belz@brighton.ac.uk

Adrian Muscat
Communications and Computer Engineering
University of Malta
Msida MSD 2080, Malta
adrian.muscat@um.edu.mt

Pierre Anguill **Mouhamadou Sow** **Gaétan Vincent** **Yassine Zinessabah**
INSA Rouen Normandie
685 Avenue de l'Université
76800 Saint-Etienne-du-Rouvray, France

Abstract

We present SpatialVOC2K, the first multilingual image dataset with spatial relation annotations and object features for image-to-text generation, built using 2,026 images from the PASCAL VOC2008 dataset. The dataset incorporates (i) the labelled object bounding boxes from VOC2008, (ii) geometrical, language and depth features for each object, and (iii) for each pair of objects in both orders, (a) the single best preposition and (b) the set of possible prepositions in the given language that describe the spatial relationship between the two objects. Compared to previous versions of the dataset, we have roughly doubled the size for French, and completely reannotated as well as increased the size of the English portion, providing single best prepositions for English for the first time. Furthermore, we have added explicit 3D depth features for objects. We are releasing our dataset for free reuse, along with evaluation tools to enable comparative evaluation.

1 Introduction

Research in image labelling, description and understanding has a long tradition, but has recently seen explosive growth. Work in this area is most commonly motivated in terms of accessibility and data management, and has a range of different specific application tasks. One current research focus is detection of relations between objects, in particular for image description generation, and the research presented here contributes to this line of work with a new dataset, SpatialVOC2K,[1] in which object pairs in images have been annotated with spatial relations encoded as sets of prepositions, specifically for image-to-text generation. We start below with the source datasets from which we obtained the images, bounding boxes, and candidate prepositions (Section 2), followed by an overview of directory structure and file schemas (Section 3), and a summary of the annotation process (Section 4) and spatially relevant features (Section 5). We describe the two evaluation tools supplied with the dataset (Section 6), and finish with a survey of other datasets with object relation annotations (Section 7).

2 Source Data

Our main data source for SpatialVOC2K was the PASCAL VOC2008 image dataset (Everingham et al., 2010) in which every object belonging to one of 20 object classes is annotated with class label, bounding box (BB), viewpoint, truncation, occlusion, and identification difficulty (Everingham et al., 2010). Of these annotations we retain just the BB geometries and the class labels (aeroplane, bird, bicycle, boat, bottle, bus, car, cat, chair, cow, dining table, dog, horse, motorbike, person, potted plant, sheep, sofa, train, tv/monitor).

We also used Rashtchian et al.'s VOC'08 1K corpus (2010), which has 5 descriptions per im-

[1] https://github.com/muskata/SpatialVOC2K

age obtained via Mechanical Turk for 50 images from each VOC2008 class, in order to determine an initial set of candidate prepositions for our annotations (for details see Section 4). Due to quality control measures, the VOC'08 1K descriptions are of relatively high quality with few errors.

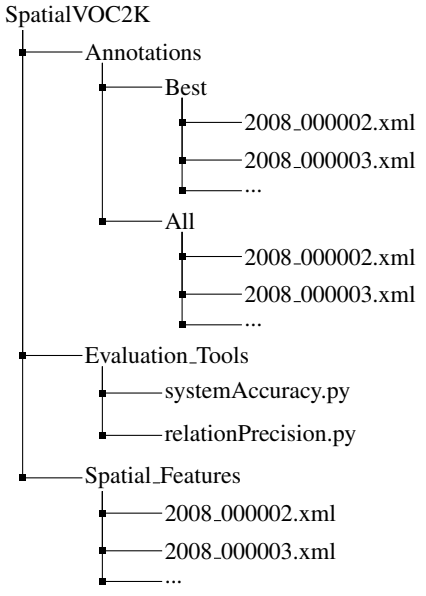
For SpatialVOC2K, we selected all images from the VOC2008 data that had two or three object bounding boxes (BBs), meaning that images contained exactly two and three objects from the VOC2008 object classes, respectively. We also selected all images with four and five BBs where three were of normal size and the remainder very small (bearing the VOC2008 label 'difficult'). This selection process resulted in a set of 2,026 images with 9,804 unique object pairs. Numbers of BBs in images were distributed as follows:

Number of BBs	2	3	3+1	3+2
Number of images	1,020	534	357	141

For each image, we then (i) collected additional annotations (Section 4) which list, for each ordered object pair, (a) *the single best*, and (b) *all possible* prepositions that correctly describe the spatial relationship between the objects; and (ii) computed a range of spatially relevant features from the image and BB geometries, BB labels, and image depth maps (Section 1). All annotations and features are included in this dataset release.

3 SpatialVOC2K Structure and Schemas

The overall structure and file conventions of the SpatialVOC2K dataset mirror those of the VOC2008 dataset where possible:



All files in the *Annotations* directory start with a line that is simply the original annotations from VOC2008. In the *Best* subdirectory, the remaining lines have the pattern $\texttt{Object}_1$ $\texttt{Object}_2$ `Preposition`, where $\texttt{Object}_1$ and $\texttt{Object}_2$ are the exact word strings, including any subscripts, of the object labels in the first line in the file, and `Preposition` is the single best preposition chosen by annotators for the two given objects presented in the given order (more about object order in Section 4 below). Each pair of annotated objects is thus associated with exactly two prepositions in the *Best* files, the best human-selected preposition for each order. The following is a simple example of a *Best* file:

```
1 VOC2012 2008_000008.jpg The VOC2008
  Database PASCAL VOC2008 flickr 500
  442 3 0 horse Left 0 1 53 87 471 420
  0 person Unspecified 1 0 158 44 289
  167 0
2 horse person under
3 person horse on
```

In the *All* directory, files have the same structure except that in the preposition lines, instead of a single preposition, there are as many prepositions as were selected by the human annotators as possible for a given ordered object pair.

The *Spatial_Features* files also have the same basic structure, except that instead of prepositions, there are 19 feature-value pairs (explained in Section 5) for each ordered object pair (some feature values differ depending on object order), e.g.:

```
1 VOC2012 2008_000008.jpg The VOC2008
  Database PASCAL VOC2008 flickr 500
  442 3 0 horse Left 0 1 53 87 471 420
  0 person Unspecified 1 0 158 44 289
  167 0
2 horse person F0 12 F1 14 F2 0.65 F3
  0.42 F4 1.54 ...
3 person horse F0 14 F1 12 F2 0.42 F3
  0.65 F4 0.65 ...
```

In the following three sections, we explain how we obtained the preposition annotations and spatial features, and how the metrics encoded by the evaluation tools are defined.

4 Preposition Annotations

We derived a set of candidate prepositions from the VOC2008 1K dataset by parsing the 5,000 descriptions in it with the Stanford Parser version 3.5.2[2] with the PCFG model, extracting the nmod:*prep* prepositional modifier relations, and

[2]http://nlp.stanford.edu/software/lex-parser.shtml#Download

manually removing the non-spatial ones. This gave us a set of 38 English prepositions:

$\mathbf{V_E^0}$ = { *about, above, across, against, along, alongside, around, at, atop, behind, below, beneath, beside, beyond, by, close to, far from, in, in front of, inside, inside of, near, next to, on, on top of, opposite, outside, outside of, over, past, through, toward, towards, under, underneath, up, upon, within* }

To obtain prepositions for French, we first asked two French native speakers to compile a list of possible translations of the English prepositions, and to check these against 200 sample images randomly selected from the complete set to be annotated. This produced 21 prepositions which were reduced to 19, based on evidence from previous work (Muscat and Belz, 2015), by eliminating prepositions that were used fewer than three times by annotators (*en haut de, parmi*). After the first batch of 1,020 images had been annotated, we furthermore merged prepositions which co-occur with another preposition more than 60%[3] of the times they occur in total (*á l'interieur de, en dessous de*), in accordance with the general sense of synonymity defined in the previous work (Muscat and Belz, 2017). We found this kind of co-occurrence to be highly imbalanced, e.g. the likelihood of seeing *á l'interieur de* given *dans* is 0.43, whereas the likelihood of seeing *dans* given *á l'interieur de* is 0.91. We take this as justification for merging *á l'interieur de* into *dans*, rather than the other way around. The process leaves a final set of 17 French prepositions:

$\mathbf{V_F}$ = {*à côté de, á l'éxterieur de, au dessus de, au niveau de, autour de, contre, dans, derrière, devant, en face de, en travers de, le long de, loin de, par delà, près de, sous, sur*}

We also reduced the set of 38 English prepositions, as follows. Following an earlier set of annotations (Belz et al., 2015), we removed the 14 prepositions that were used five times or fewer by annotators; next, for each of the 3 strongest synonym sets (2 other sets coincided with changes made because of the frequency threshold), we retained only the single preposition most frequently used by the annotators, and overwrote the other members of the set with it, yielding a final set of 20 English prepositions:

$\mathbf{V_E}$ = { *above, against, alongside, around, behind, below, beneath, beside, beyond, by, close to, far from, in, in front of, near, next to, on, on top of, opposite, under* }

As discussed in more detail in previous work (Muscat and Belz, 2017), we make the domain-specific assumption that there is a one-to-one mapping from each preposition to the SR it denotes (whereas an SR can map to multiple prepositions). While our machine learning task is SR detection, we ask annotators to annotate our data with the corresponding prepositions (a more human-friendly task).

We used the above preposition sets in collecting annotations as follows. For each object pair O_i and O_j in each image, and for both orderings of the object labels, L_i, L_j and L_j, L_i, the task for annotators was to select (i) the single best preposition for the given pair (free text entry), and (ii) the possible prepositions for the given pair (selected from a given list) that accurately described the relationship between the two objects in the pair, given the template L_1 *is* ___ L_2 (*is* becomes *et* for French).

Even though in annotation task 1, annotators were not limited in their choice of preposition, they did not use any that were not in the list of prepositions offered in annotation task 2 (a few typos we corrected manually). As it would have been virtually impossible to remember the exact list of prepositions and only use those, we interpret this as meaning that annotators did not feel other prepositions were needed.

We used average pairwise kappa to assess inter-annotator and intra-annotator agreement as described in previous work (Muscat and Belz, 2017). First, figures for the first batch of French annotations (1,020 images with 2 or 3 objects in BBs[4]). For *single best* prepositions (annotation task 1), average inter-annotator agreement was 0.67, and average intra-annotator agreement was 0.81. For *all possible* prepositions (annotation task 2), average inter-annotator agreement was 0.63, and average intra-annotator agreement was 0.77.

For the second batch of French annotations (1,006 images with 3, 4 or 5 BBs), average inter-annotator agreement for *single best* prepositions (annotation task 1) was 0.33, and average intra-annotator agreement was 0.66. For *all possible* prepositions (annotation task 2), average inter-

[3] This is a very high threshold and far above co-occurrence percentages for any other preposition pairs.

[4] Annotators were only ever shown images with 2 BBs in them.

$F0$:	Object label L_s — definition depends on learning method	NB, DT, RF: {0, 1, ..., 19}; others: 1-hot encoding (20 bits)
$F1$:	Object label L_o — definition depends on learning method	
$F2$:	Area of bounding box of Obj_s normalized by image size.	[0, 1]
$F3$:	Area of bounding box of Obj_o normalized by image size.	[0, 1]
$F4$:	Ratio of Obj_s bounding box area to that of Obj_o.	[0, size of Obj_s]
$F5$:	Distance between bounding box centroids, normalized by image diagonal.	[0, 1]
$F6$:	Area of overlap of bounding boxes normalized by the area of the smaller bounding box.	[0, 1]
$F7$:	Distance between centroids divided by sum of square root of areas/2 (approximated average width of bounding boxes).	[0, ∼20]
$F8$:	Position of Obj_s relative to Obj_o expressed as one of 4 categories, depending on the angle with the vertical axis.	NB, DT, RF: {0, 1, 2, 3}; others: 1-hot encoding (4 bits)
$F9$–$F12$:	Let distance from image edge of left and right edges be $a1, b1$ for first box and $a2, b2$ for second box: $F9 = (a2-a1)/(b1-a1)$, $F10 = (b2-a1)/(b1-a1)$. Similarly for the top and bottom edges, giving $F11$ and $F12$.	[∼-40, ∼+40]
$F13$:	Aspect ratio of box of Obj_s.	[0, ∼10]
$F14$:	Aspect ratio of box of Obj_o.	
$F15$:	GloVe word vector for L_s.	here: $\sim [-2, +3]$
$F16$:	GloVe word vector for L_o.	
$F17$:	Average depth in BB of Obj_s.	
$F18$:	Average depth in BB of Obj_o.	

Table 1: Spatially relevant features as included in SpatialVOC2K. Note that the 19 numbered features above correspond to feature vectors of length between 116 and 140, depending on conversion method for ML inputs.

annotator agreement was 0.3, and average intra-annotator agreement was 0.62. A possible reason for the lower annotator agreement on batch 2 is that as the number of dominant objects in an image increases, the annotation task becomes more difficult; we also used different annotators for the second batch which may be a contributing factor.[5]

5 Spatially Relevant Features

Table 1 provides an overview of the 19 features included in SpatialVOC2K: F0, F1, F15 and F16 are language features. F0 is the class label of the first object, F1 of the second (e.g. *person*). F15 and F16 are GloVe word vectors of length 50 (Pennington et al., 2014) for the object labels.[6] F2–F14 are visual features measuring various aspects of the geometries of the image and two bounding boxes (BBs). Most features express a property of just one of the objects, but F4–F9 express a property of both objects jointly, e.g. F6 is the normalized BB overlap.

F17 and F18 are the average pixel-level depth value within the BB of Obj_s and Obj_o, respectively. Pixel-level depth values were computed via the method described in (Birmingham et al., 2018), which uses depth maps computed with monoDepth[7] (Godard et al., 2017) .

6 Evaluation Tools

SpatialVOC2K includes two evaluation tools which we have used in all previous work involving similar data. The two tools, `systemAccuracy` and `relationPrecision` implement the following two methods, respectively.

System-level Accuracy: There are four different variants of system-level Accuracy, denoted $Acc(n)$, $n \in \{1, 2, 3, 4\}$. Each variant returns Accuracy rates for the top n outputs returned by systems, in the sense that a system output is considered correct if at least one of the reference prepositions (the human-selected prepositions from the dataset annotations) can be found in the top n prepositions returned by the system (for $n = 1$ this yields standard Accuracy).

Weighted Average Per-preposition Precision: This measure, denoted Acc_P, computes the weighted mean of individual per-preposition precision scores. The individual per-preposition precision for a given system and a given preposition p is the proportion of times that p is among the corresponding human-selected prepositions out of all the times that p is returned as the top-ranked preposition by the system.

[5] Inter-AA/intra-AA for English and additional dataset statistics will be added to the project home on Github.

[6] GloVe is a count-based method for creating distributed word representations.

[7] https://github.com/mrharicot/monodepth

Name	Authors	Task	Categories of relations	Annotated relations	Images
Visual Phrases	Sadeghi et al. 2011	Phrase Classification	action, verbal, spatial	1,796	2,769
Visual and Linguistic Treebank	Elliott and Keller, 2013	Image Description	action, verbal, spatial	5748	341 / 2424
Scene Graphs	Johnson et al. 2015	Image Retrieval	action, verbal, spatial, preposition	112,707	5K
ViSen	Ramisa et al. 2015	Preposition Prediction	spatial, preposition	78,317	33,262
VRD	Lu et al. 2016	Relation, Phrase Prediction	action, verbal, spatial, preposition, comparative	37,993	5K
Visual Genome	Krishna et al. 2016	Image Understanding	action, verbal, spatial, preposition, comparative	1.5M	108K

Table 2: Overview of related datasets. For explanation of relation categories see in text.

7 Related Datasets

A number of datasets are available that incorporate annotations representing relations between objects in images. Types of relationships that have been annotated include actions (e.g. *person kicks ball*), other verbal relations (*person wears shirt*), spatial relations (*person on horse*), and comparative relations (*one car bigger than another*). In this section, we provide a brief overview of available datasets with relation annotations, in terms of their stated purpose (application task), the types of relations included, the range of spatial prepositions included, as well as size and other properties of the dataset. Table 2 has a summary of the datasets.

Visual Phrases (Sadeghi and Farhadi, 2011) was the first image dataset with object relation annotations, and used the concept of a visual phrase (VP) which is defined as a bounding box that surrounds two objects in an image. Out of 17 different types of VPs annotated in the data set, 13 comprise 2 objects, and 4 comprise one object. However, there are 120 predicates per object category.

Visual and Linguistic Treebank (Elliott and Keller, 2013) contains 341 images that are annotated with regions (362 in total) and visual dependency representations, which unfold to a total of 5,748 spatial relations (from a set of 8) and are aligned to the dependency parse of the image description. This setup allows for the prediction of actions as well as spatial relations (using a set of 8 manual created rules).

Scene Graphs (Johnson et al., 2015) is a dataset of 5,000 human-generated scene graphs grounded to images; scene graphs describe objects and their relationships.

ViSen (Ramisa et al., 2015) associates sets of (object_1, preposition, object_2) triples with images, where the triples have been extracted from parses of the image descriptions in MSCOCO (Lin et al., 2014) and Flickr30k (Young et al., 2014). Prepositions covered include all those extracted from the image descriptions including non-spatial ones. By far not all descriptions contain prepositions so not all images have spatial relation annotations; the task addressed is preposition prediction, not spatial relation prediction.

Visual Relationships Dataset (VRD) (Lu et al., 2016) contains 5,000 images, 100 object categories, 6,672 unique relationships, and 24.25 relations per object category. Scant information is available about how the dataset was created other than that relations broadly fit into the categories action, verbal, spatial, preposition and comparative.

Visual Genome (Krishna et al., 2017) contains 108K images, split into 4M regions, corresponding to 108K scene graphs and about 4K region graphs, 1.5M object-object relations, 40K unique relations, and an average of 17 relations per image and 0.63 relations per region.

8 Future Work

We plan to expand the SpatialVOC2K dataset to other languages, and to more object pairs per language, in the future. Given the ever growing need for image description and labelling, and in combination with the image segmentation and description annotations that exist for the same VOC images, SpatialVOC2K can potentially be used in a range of different application tasks, including but not limited to image description generation.

References

A. Belz, A. Muscat, M. Aberton, and S. Benjelloun. 2015. Describing spatial relationships between objects in images in english and french. In *4th Workshop on Vision and Language*, pages 104–113, Lisbon, Portugal.

B. Birmingham, A. Belz, and A. Muscat. 2018. Adding the third dimension to spatial relation detection in 2d images. In *Proceedings of INLG'18*.

D. Elliott and F. Keller. 2013. Image description using visual dependency representations. In *Proc. 18th Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1292–1302, Seattle.

M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. 2010. The pascal visual object classes (voc) challenge. *Int. J. of Computer Vision*, 88(2):303–338.

C. Godard, O. M. Aodha, and G. J. Brostow. 2017. Unsupervised monocular depth estimation with left-right consistency. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6602–6611.

J. Johnson, R. Krishna, M. Stark, L. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei. 2015. Image retrieval using scene graphs. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3668–3678.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, pages 1–42.

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*, pages 740–755. Springer.

Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2016. Visual relationship detection with language priors. In *Computer Vision – ECCV 2016*, pages 852–869, Cham. Springer International Publishing.

A. Muscat and A. Belz. 2015. Generating descriptions of spatial relations between objects in images. In *Proc. 15th European Workshop on Natural Language Generation (ENLG)*, pages 100–104, Brighton, UK.

Adrian Muscat and Anja Belz. 2017. Learning to generate descriptions of visual data anchored in spatial relations. *IEEE Computational Intelligence Magazine*, 12(3):29–42.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proc. 19th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar.

Arnau Ramisa, Josiah Wang, Ying Lu, Emmanuel Dellandrea, Francesc Moreno-Noguer, and Robert Gaizauskas. 2015. Combining geometric, textual and visual features for predicting prepositions in image descriptions. In *Proc. 20th Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pages 214–220, Lisbon, Portugal.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using amazon's mechanical turk. In *Proc. NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147, Los Angeles, California.

Mohammad Amin Sadeghi and Ali Farhadi. 2011. Recognition using visual phrases. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1745–1752. IEEE Computer Society.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.