Combining Lexical and Spatial Knowledge to Predict Spatial Relations between Objects in Images

Manuela Hürlimann
The Insight Centre for Data Analytics
National University of Ireland Galway
IDA Business Park, Lower Dangan, Galway
manuela.huerlimann@insight-centre.org

Johan Bos
Center for Language and Cognition
University of Groningen
The Netherlands
johan.bos@rug.nl

Abstract

Explicit representations of images are useful for linguistic applications related to images. We design a representation based on first-order models that capture the objects present in an image as well as their spatial relations. We take a supervised learning approach to the spatial relation classification problem and study the effects of spatial and lexical information on prediction performance. We find that lexical information is required to accurately predict spatial relations when combined with location information, achieving an F-score of 0.80, compared to a most-frequent-class baseline of 0.62.

1 Introduction

In the light of growing amount of digital image data, methods for automatically linking data to language are a great asset. Due to recent advances in the distinct areas of language technology and computer vision, research combining the two fields has become increasingly popular, including automatical generation of captions (Karpathy and Fei-Fei, 2014, Elliott and Keller, 2013, Elliott et al., 2014, Kulkarni et al., 2011, Vinyals et al., 2014, Yang et al., 2011) and translation of text into visual scenes (Coyne et al., 2010).

One task which has not yet been extensively researched is the automatic derivation of rich abstract representations from images (Neumann and Möller, 2008, Malinowski and Fritz, 2014). A formal representation of an image goes beyond naming the objects that are present; it can also account for some of the *structure* of the visual scene by including spatial relations between objects. This information could enhance the interface between language and vision. Imagine, for instance, searching for images that show a "man riding a bicycle": it is necessary, but not sufficient, for pictures to contain both a man and a bicycle. In order to satisfy the query, the man also has to be somehow connected to the bicycle, with his feet on the pedals and his hands on the steering bar.

We argue that representations of images which take into account spatial relations can enable more sophisticated interactions between language and vision that go beyond basic object co-occurrence. The aim of this paper is to use an extension of first-order models to represent images of real situations. In order to obtain such models, we need (a) high-quality, broad-coverage object localisation and identification and methods to (b) accurately determine object characteristics and to (c) detect spatial relationships between objects.

As broad-coverage object detection systems are not yet available, we carry out steps (a) and (b) manually. Hence, in this paper, we focus on step (c): the detection of spatial relations. This is difficult because there is a vast number of ways in which a given relation can be realised in a visual scene. The questions that we want to answer are whether first-order models of classical logic are appropriate to represent images, and what features are suitable for detecting spatial relationships between objects in images. In particular, we want to investigate what the impact of lexical knowledge is on determining spatial relations, independent of the quality of object recognition.

This paper is organised as follows. We will first give more background about spatial relations (Section 2) and related work on combining vision with language technology (Section 3). Then we will introduce our data set in Section 4, comprising a hundred images with a total of 583 located objects for which spatial relations need to be determined. In Section 5 we outline our classification method in detail and present and discuss our results.

2 Background: Spatial Relations

In this paper we focus on the task of predicting spatial relations in images, investigating three relations (`part-of`, `touching`, `supports`; see Section 4). We integrate the detected spatial relations into first-order models borrowed from logic, which offer an easily extendable representation of an image. Once detected, spatial relations can also serve as a useful basis for predicting more specific predicates which hold between objects, such as actions. For example, "ride" presupposes `touching`, and "carry" or "hold" presuppose that the object being carried or held is `supported` by the other object. The spatial configuration of two objects restricts the spatial relations which are possible (and plausible) between them; for example, two objects can only `touch` if they are in sufficient proximity to each other. Knowledge of objects properties further constrains the set of plausible relations. For example, if asked to determine whether the two objects in Figure 1 are in a `part-of` relationship, the decision is difficult on spatial grounds alone, that is, not knowing *what* objects are (indicated by blackening the picture). In this case, the spatial configuration on its own does not supply sufficient information to confidently answer this question.

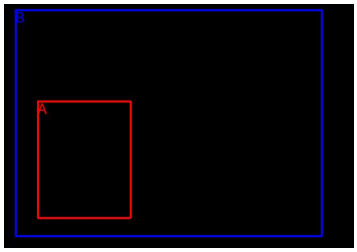

Figure 1: Is A (red) part of B (blue)? We can't tell: we need semantic knowledge of A and B.

However, information about the objects themselves, beyond their locations, improves spatial relation prediction. Consider Figure 2: when we reveal the object identities, we can be very certain that the ice cream and boy are *not* in a `part-of` relationship, but the cat and head are. Such inferences about spatial relations are straightforward for humans, while this is a difficult task for computers. We suggest, however, that useful machine-readable world knowledge can be gleaned from lexical resources such as WordNet (Miller, 1995) and large text corpora.

Figure 2: A *not* part of B (left); A part of B (right)

While many researchers have focused on generating textual descriptions for images (Karpathy and Fei-Fei, 2014, Elliott and Keller, 2013, Elliott et al., 2014, Kulkarni et al., 2011, Vinyals et al., 2014, Yang et al., 2011), deriving a first-order semantic model from an image is a task hitherto unattempted. The advantage of having an abstract model instead of a textual label is the ease with which *inferences* can be made. Inference processes include querying the model and checking for consistency and informativeness. This greatly facilitates maintenance of image databases and enables applications such as question answering and image retrieval (Elliott et al., 2014).

3 Related Work

Research into combining Natural Language Processing and Computer Vision has become increasingly popular over the past years. There is an extensive body of work, among others in the following areas: building multimodal models of meaning which take into account both text and image data (Bruni et al., 2012), generating images from textual data (Lazaridou et al., 2015, Coyne et al., 2010), Question Answering on images (Malinowski and Fritz, 2014), and automatic image label generation (Karpathy and Fei-Fei, 2014, Elliott and Keller, 2013, Elliott et al., 2014, Kulkarni et al., 2011, Vinyals et al., 2014, Yang et al., 2011).

Belz et al. (2015) present a method for selecting prepositions to describe spatial relationships between objects in images. They use features based on geometrical configurations of bounding boxes as well as prior probabilities of prepositions occurring with objects/class labels.

Several approaches have been proposed to reason on spatial information derived from visual input. Neumann and Möller (2008) discuss the potential of knowledge representation for high-level scene interpretation. Their focus is on Description Logic (DL), a subset of first-order predicate calculus supporting inferences about various aspects of the scene. They identify requirements and

processes for a system conducting stepwise inferences about concepts in a scene. This would make use of low-level visual and contextual information, spatial constraints, as well as taxonomic and compositional links between objects. As their work is a conceptual exploration of the area, they do not specify how they would acquire such a knowledge base with information about object relations and contexts.

Falomir et al. (2011) aim at creating a qualitative description of a scene (image or video still) and translating it into Description Logic. Object characteristics of interest include shape and colour as well as spatial relations. The latter are based on topology and include *disjoint*, *touching*, *completely inside*, and *container* as well as information about relative orientation of objects. All qualitative descriptions are aggregated into an ontology with a shared vocabulary, which aids the inference of new knowledge using reasoning.

Zhu et al. (2014) present a Knowledge Base (KB) approach to predicting affordances (possibilities of interacting with objects - e.g. the handle on a teacup is an affordance for holding). Evidence in their Markov Logic Network KB consists of: affordances (actions), human poses, five relative spatial locations of objects with respect to the human (*above*, *in-hand*, *on-top*, *below*, *next-to*), and the following kinds of attributes: visual (material, shape, etc; obtained using a visual attribute classifier), physical (weight, size; obtained from online shopping sites), and categorical (hypernym information from WordNet). They stress the importance of inference, which is an essential benefit of their approach. Their results for zero-shot affordance prediction show a clear improvement compared to classifier-based approaches, underlining the strength of the KB approach. They find that categorical ("lexical") attributes boost performance.

4 The Image Model Collection

Below we present GrImSem-100 (Groningen Image Semantics - 100), the dataset used in the present work, which comprises a set of images paired with image models. The image models contain the first-order objects present in the images together with their spatial relations. First we describe the selected images and how we annotated them with spatial relations. Then we show what kind of models we use to represent the images.

4.1 Selected Images

Our dataset consists of one hundred images with associated first-order semantic models. We carefully hand-picked copyright-free images from an existing large image resource.[1] The selected images are shown in Figure 3. In the image selection

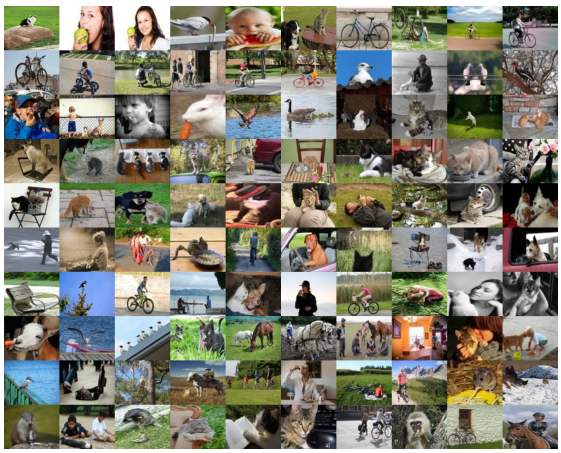

Figure 3: Selected images of our corpus.

process only images were chosen that contained two or more clearly visible concrete real-world objects, in order to get image material interesting for investigating spatial relation between various objects. As a result, typical images are of dogs chasing cats, human beings or animals eating something, or people riding their bicycle.

Selection of objects to annotate was mostly based on object size (large objects are annotated, small ones omitted), but exceptions were made for small objects which were striking or interesting. Each object was captured by a *bounding box*, also known as a "Minimal Bounding Rectangle" (MBR), a often used approximation to identify object in images (Wang, 2003). The bounding box of an object (Figure 4) is simply a rectangle covering all of its extent, thus preserving the object's "position and extension" (Wang, 2003). In total, 583 objects from 139 different synset categories were annotated across the 100 images.

4.2 Spatial Relations

In the scope of this paper we investigated three spatial relations:

- `part-of`

[1] Pixabay, `https://pixabay.com/en/`. All images are free to use, modify and distribute under the Creative Commons Public Domain Deed CC0 `https://creativecommons.org/publicdomain/zero/1.0/`, for both commercial and academic purposes

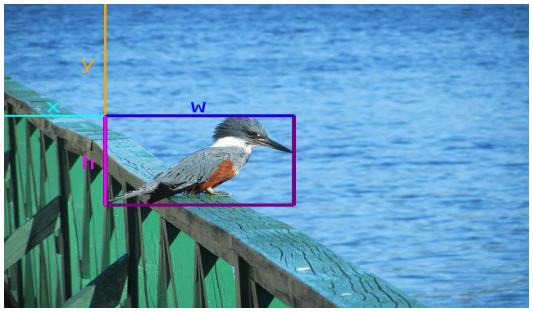

Figure 4: Bounding boxes with coordinates.

- touching
- supports

We selected part of, touching and supports for prediction because they are well-defined and less fuzzy than for example "far" or "near" / "close". Part of is closely connected to the part meronymy relation from lexical semantics and therefore interesting for our approach, which uses lexical knowledge. Touches and supports can be considered useful for predicting further predicates, such as actions. Additionally, we annotated a fourth spatial relation in the models, occludes, because we thought it would be an important feature in predicting the other three spatial relations. Below we discuss the properties of each of these relations.

Part-of If object A is part-of object B, then A and B form an entity such that if we removed A, B would not be the same entity any more and could not function in the usual way (e.g. A - wheel, B - bicycle). The part-of relation is transitive and asymmetric. Furthermore, no object can be part-of itself.

Touching Two objects A and B are touching if they have at least one point in common; they are not disjoint. Only solid and fluid, but not gaseous objects (such as "sky") can be in a touching relation. Touching is always symmetric but not necessarily transitive.

Supports In order for object A to support object B, the two objects need to be touching. Support means that the position of A depends on B: if B was not there, A would be in a different position. Therefore, there is the notion of "support against gravity", discussed by Sjöö et al. (2012, p.8). Supports can be mutual (symmetric), but this is not a requirement; in fact, asymmetric support is probably more frequent. Furthermore, supports is transitive. For example, if a table supports a plate, and the plate supports a piece of cake, then the table also supports the piece of cake.

Occludes If object A occludes object B, it renders it partly invisible. Occlusion is viewpoint-sensitive: from the point of view of the observer, object A is partly in front of object B. For example, in Figure 6, the cat occludes the armchair.

4.3 Annotating Spatial Relations

We used the Crowdflower crowdsourcing platform to annotate the gold standard for the three spatial relations. In all annotation tasks, workers were presented with an image which had two objects highlighted in bounding boxes (one red and one blue). They had to choose the statement which they deemed to best describe the relationship between the two objects. To facilitate identification of objects in cluttered pictures, we provided the first WordNet lemma of the synset as a label for each box, prefixing with "A" and "B" for the directed relations part-of and supports. Figure 5 shows an example question as presented in the part-of task.

Figure 5: Example question presented to Crowdflower workers on part-of task.

Post-processing of the raw annotation results was done using the Multi-Annotator Confidence Estimation tool, MACE (Hovy et al., 2013). MACE is designed to evaluate data from categorical multi-annotator tasks. It provides competence ratings for individual annotators as well as the most probable answer for each item. A subsample of the MACE output was assessed manually and errors found during this inspection were corrected. However, a little bit of noise is likely to remain in the final spatial relation annotations.

4.4 Image Models and Grounding

In classical logic, a first-order model $M = \langle D, F \rangle$ has two components, a non-empty *domain* D (also called *universe*) and an *interpretation function* F (Blackburn and Bos, 2005). The domain is the set of all entities occurring in the model, and the interpretation function maps non-logical symbols from the vocabulary to these entities. We adopt the Prolog-readable model format of Blackburn & Bos for for our set of 100 images.

Each image is thus paired with a model that describes its key features, providing a simplified representation of the reality depicted in the image. The *vocabulary* of non-logical symbols present in the models is based on WordNet (Miller, 1995): we use the names of noun *synsets* as one-place predicates to name entities, and those of adjectives for modelling attributes (such as colours). Hyperonyms from a pruned top-level ontology were also semi-automatically added to the model to further enrich the image models. Additionally, we introduce two-place relations for the four spatial relations introduced in the previous section: `s_part_of`, `s_touch`, `s_supports`, and `s_occludes`.

Since we also model spatial characteristics of the situations at hand, we need to be able to ground the entities in the model to its physical location in the image. We do this with the help of a *grounding* function `G`. As a consequence, our grounded first-order models are defined as $M = \langle D, F, G \rangle$. The grounding function maps the domain entities to their coordinates, that is, the location in pixel space represented by bounding boxes. For the coordinates, we use the Pascal VOC notation (Everingham and Winn, 2012, p. 13), as illustrated in Figure 4. All distances are measured in pixels. An example of a model including Domain `D`, Interpretation Function `F` and Grounding `G` can be seen in Figure 6.

5 Predicting Spatial Relations

5.1 Instances

Based on our image-model dataset (see Section 4), we create a set of object pairs for classification purposes. All ordered combinations of two objects (pairs) within an image are considered, giving us a total of 1,515 instances for classification. We randomly split the instances (across all images), using 90% (1,364 pairs) for training purposes and reserving 10% (151 pairs) as unseen test data.

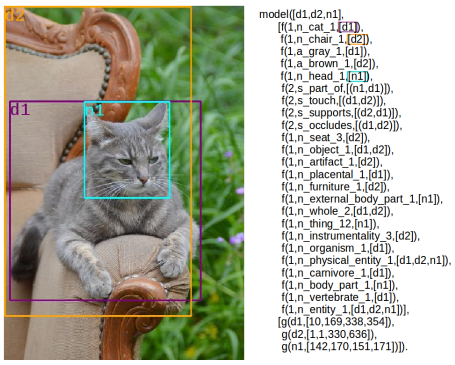

Figure 6: Image and grounded first-order model.

Table 1: Distribution of class labels in training and testing data.

relation	training	testing	overall
A part of B	16	2	18
B part of A	148	16	164
A and B touch	137	16	153
A and B touch + A supports B	86	9	95
A and B touch + B supports A	119	14	133
no relation	858	94	952
total	1,364	151	1,515

5.2 Task Formulations

We cast the spatial relation prediction task as a classification problem, in which each instance belongs to one of the following disjoint classes:

- A part of B
- B part of A
- A and B touch
- A and B touch + A supports B
- A and B touch + B supports A
- no relation: A and B are in no relation

Table 1 shows the distribution of the classes across the training and testing (unseen) data.
We distinguish two subtasks according to the set of instances selected for classification:

- **Subtask A:** predicting relation existence and types (all instances)
- **Subtask B:** predicting relation types only (excluding the class "no relation")

We use a multi-label formulation, i.e. the labels *A part of B*, *B part of A*, *touching*, *A supports B* and *B supports A* are used, and each instance can have multiple labels (or none).

5.3 Features

5.3.1 Spatial Features

The spatial features capture knowledge about the spatial properties of (pairs of) objects.

Overlap This consists of two features:

- a boolean: do the two bounding boxes have at least one pixel in common?
- the size of this overlap, that is, the number of pixels that the two bounding boxes share

Contained-in Two booleans expressing whether (i) the bounding box of the first object is entirely contained within that of the second object or (ii) vice versa.

Object size We approximate true size by using the surface area of the corresponding bounding box (in pixels). In order to account for the effects of object truncation, varying image sizes and perspectives, we average in two steps for each synset. First, we normalise the size (width x height) of each object in each image by the width and height of the image. Second, we average these normalised surface areas for each object type (e.g. `cat.n.01`) across all images, obtaining the following features:

- The size of the first object
- The size of the second object
- The absolute difference in size between the first and the second object

Occlusion Occlusion carries information about the depth alignment of objects. An object occludes another if it partially renders it invisible (see Section 4.2). CrowdFlower was used to annotate occlusion (see Section 4.3).

5.3.2 Lexical Features

Lexical features capture linguistics knowledge about objects from WordNet and corpora.

Meronymy (part-whole relation) For a pair of objects (A, B) we determine whether A is a part meronym of B, or B is a part meronym of A (two boolean features).

Hypernymy In addition to information about meronyny (`has-a`), we also consider the ontological `is-a` status of objects. We use a top-level pruned ontology, which is divided into ten levels, to obtain the following features for each level (Blanchard et al., 2005):

1. Are the hypernyms identical? (boolean)
2. Path similarity of the hypernyms (range 0-1)
3. Leacock-Chodorow (LCH) similarity (no fixed range)
4. Wu-Palmer (WUP) similarity (no fixed range)

Corpus features Useful information about objects can be gleaned from large text collections. We thus use co-occurrence data from the first ten subcorpora of the ukWaC corpus comprising 92.5 million words (Baroni et al., 2009).

For each instance, we extract all uni-, bi- and trigrams (excluding sentence-final punctuation) that occur between lemmas of the first and lemmas of the second object. From these data, we extract the following feature sub-groups:

1. prepositions (pos-tag `IN`) - e.g. "cat *on* (the) lawn"
2. verb forms of "to have" and "to be" (pos-tags `VH.?` and `VB.?`)
3. verb forms of other verbs (pos-tag `VV.?`)

We consider single prepositions and verbs as well as sequences of two prepositions or two verbs. The raw data for prepositions and "other" verbs are reduced according to greatest coverage, retaining 50 and 100, respectively. In classification, for an ordered pair of objects, we use the *frequency* with which the given verb or preposition occurs across all lemma pairs as a feature.

Word embeddings Word embeddings are another way to make use of co-occurrence data. We use the pre-trained 300-dimensional `word2vec` vectors by Mikolov et al. (2013a) and Mikolov et al. (2013b). These vectors were trained on a 100 billion-word subpart of the Google News dataset. We calculate the vector for each synset as an average across the vectors of all its lemmas. In order to obtain features from a pair of synsets the second vector is subtracted from the first and each dimension of the resulting vector is added as a feature (300 features).

5.4 Results

We evaluate prediction performance using the F1-score, obtained using 5-fold stratified cross-validation and averaged across two runs. We report scores for each relation as well as micro-averaged overall scores.

Combo1	configuration — bounding box overlap, contained in, occlusion (6 features)
Combo2	size (3 features)
Combo3	meronymy (2 features)
Combo4	hypernym identity (10 features)
Combo5	hypernym similarity measures (30 features)
Combo6	co-occurrence frequency with prepositions (50 features)
Combo7	word embedding subtraction (300 features)
Combo8	co-occurrence frequency with verbs other than "to have" and "to be" (100 features)
Combo9	co-occurrence frequency with "to have" and "to be" (7 features)

Table 2: Feature combinations.

A baseline choosing the most frequent label(s) would assign "no relation" in subtask A (achieving 0.623), and `touching` (without an additional `supports` label) in subtask B (achieving 0.405).

Another point of comparison is the work by Rosman and Ramamoorthy (2011). They use a data-driven contact-point approach to classify 132 instances into three different relations. They achieve an overall F-score of 0.72, with results for individual relations ranging between 0.47 to 0.84.[2]

In order to assess the effect of the spatial and lexical features, we divide the features up into the groups shown in Table 2 (Combo1 and Combo2 are spatial features, while Combo3-9 are lexical features).

We test all possible combinations without replacement of the nine groups in the range 1 to 9, separately on (i) the set of all instances (subtask A) and on (ii) the set of instances which are in a relation (subtask B). In order to evaluate the results, we calculate the average F-score for each single feature group (1, 2, 3, ...) as well as for combinations of feature groups (1+2, 1+2+3, 1+2+3+4, ...). There are 511 possible combinations.

In Table 3 we report the baselines, the best single groups (Combo3 (meronymy) in subtask A; Combo1 (spatial configuration in subtask B), spatial groups only, lexical groups only and the best respective combinations per subtask (1+2+3+5 in subtask A; 1+2+3+9 in subtask B). A number of interesting things can be observed: first, all approaches significantly outperform the baselines if we combine multiple groups of features. Second,

[2]These figures were calculated from the confusion matrix in Rosman and Ramamoorthy (2011, p. 16).

	subtask A	**subtask B**
baseline	0.62	0.41
single groups	0.71[a]	0.74[b]
only spatial (groups 1 + 2)	0.78	0.82
only lexical (groups 3-9)	0.68	0.72
best lexical+spatial	**0.80**[c]	**0.85**[d]

[a]Group 3 (meronymy), best single group in subtask A
[b]Group 1 (spatial configuration), best single group in subtask B
[c]1+2+3+5, best combination in subtask A
[d]1+2+3+9, best combination in subtask B

Table 3: Summary of results on training data (overall F-scores).

	subtask A	**subtask B**
baseline	0.62	0.41
single groups	0.65	0.72
only spatial (groups 1 + 2)	0.80	0.80
only lexical (groups 3-9)	0.66	0.69
best lexical+spatial	**0.82**	**0.86**

Table 4: Summary of results on unseen test data (overall F-scores).

performance on subtask B is generally better than on subtask A, indicating that pre-selecting object pairs which are in a relation facilitates prediction. Third, the combined spatial feature groups perform better than the combined lexical feature groups; however, the best models are those which *combine* features from the spatial and lexical domain. Experiments on the reserved test set (see Table 4) further confirm that overfitting is not an issue and that the results obtained using cross-validation are robust.

Looking at performance for the individual relations, we find that `part-of` yields the best results[3] (achieving F-scores of 0.95 in subtask A and 0.96 in subtask B), while `touching` is the most difficult to predict (0.48 in subtask A - below baseline; 0.76 in subtask B). For `supports` we achieve 0.71 on subtask A and 0.88 on subtask B, and no relation (only in subtask A) scores 0.88.

5.5 Error Analysis

Tables 5 and 6 show the confusion matrices for the respective best-performing combinations of feature groups. Generally, it is straightforward to identify the direction of a relation, that is, to distinguish between `A part of B` and

[3]F-scores mentioned are from classification optimised for individual relations.

true \ assigned	A part of B	B part of A	touching	A supports B	B supports A	no relation
A part of B	**13**	0	0	0	0	3
B part of A	10	**139**	4	0	0	5
touching	10	6	**73**	1	5	52
A supports B	10	0	22	**43**	1	20
B supports A	10	0	21	1	**66**	31
no relation	10	3	67	6	17	**765**

Table 5: Confusion matrix for subtask A, using feature groups 1, 2, 3 and 5.

true \ assigned	A part of B	B part of A	touching	A supports B	B supports A	no relation
A part of B	**14**	0	2	0	0	0
B part of A	1	**143**	4	0	0	1
touching	1	8	**114**	3	10	2
A supports B	1	0	17	**65**	3	1
B supports A	1	0	26	2	**91**	0

Table 6: Confusion matrix for subtask B, using feature groups 1, 2, 3 and 9.

`B part of A` and between `A supports B` and `B supports A`. We can see from Table 5 that instances which are in "no relation" (the majority class) can be identified rather unambiguously, and also the distinction between `part-of` versus `touching` / `supports` can be easily made. However, there is considerable confusion between `touching` and `support`, which are fairly frequently confused for each other, as well as for "no relation", if present. The distinction between `touching` and "no relation" is presumably due to the incidental nature of the former (`touching` strongly dependets on the local spatial configurations, but can be ambiguous / difficult to see). Pixel-level features could help improve discrimination for these. `touching` and `supports` are difficult to distinguish because they are very similar. Since `supports` is misclassified as `touching` much more often than vice versa, more discriminative features for the former need to be found in order to resolve this issue. These could address object properties such as mass/weight, but also a refinement of the prepositional features already implemented could help, for example association measures such as Mutual Information instead of the simple co-occurrence frequencies used in the present system.

6 Conclusions and Future Work

First-order models, as used in classical logic, are suitable for representing images in an abstract way. The entities in a model can be mapped to non-logical symbols from an existing ontology (we used WordNet in this paper). Spatial relations between entities can be simply added to the models. The models can be simply extended with a function that maps entities to the coordinates of the bounding boxes in images.

We developed a corpus of images depicting real situations with their first-order models, effectively linking visual scenes to language. Some of the aspects involved in this process were carried out manually, such as recognizing objects in an image, but it is not unthinkable that in the future software components could fulfil this task. We trained a classifier for recognising spatial relations between objects, and what we learn is that linguistic information is required to accurately predict these relations when combined with location information. The best performance (F-scores of 0.81 and 0.85 for subtasks A and B, respectively) was obtained when combining spatial and lexical feature groups, significantly outperforming either spatial or lexical features on their own.

The corpus of images paired with spatial models that arose from this work could be used for various research topics in the future. Currently the corpus is being extended to include more images and more spatial relations. One of the relations that we are currently investigating is the vague spatial relation `near`. The corpus also contains human-generated false and true descriptions with respect to the images. In the future we want to find out whether image models as proposed in this paper are helpful to verify the truth of a statement with respect to an image.

Acknowledgments

The first author was supported by the Erasmus Mundus Programme in Language and Communication Technologies (EM LCT) as well as the SSIX Horizon 2020 project (grant agreement No 645425). Thanks to the Computational Semantics class in Groningen for supplying the initial models. We would like to thank Benno Weck and the anonymous reviewers for their helpful comments.

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Anja Belz, Adrian Muscat, Maxime Aberton, Sami Benjelloun, and INSA Rouen. 2015. Describing spatial relationships between objects in images in english and french. In *Proceedings of the 2015 Workshop on Vision and Language (VL'15)*, pages 104–113. Association for Comptuational Linguistics.

Patrick Blackburn and Johan Bos. 2005. *Representation and Inference for Natural Language. A First Course in Computational Semantics*. CSLI Publications.

Emmanuel Blanchard, Mounira Harzallah, Henri Briand, and Pascale Kuntz. 2005. A typology of ontology-based semantic measures. In *EMOI-INTEROP*.

Elia Bruni, Jasper Uijlings, Marco Baroni, and Nicu Sebe. 2012. Distributional semantics with eyes: Using image analysis to improve computational representations of word meaning. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1219–1228. ACM.

Bob Coyne, Richard Sproat, and Julia Hirschberg. 2010. Spatial relations in text-to-scene conversion. In *Computational Models of Spatial Language Interpretation, Workshop at Spatial Cognition*. Citeseer.

Desmond Elliott and Frank Keller. 2013. Image description using visual dependency representations. In *EMNLP*, pages 1292–1302.

Desmond Elliott, Victor Lavrenko, and Frank Keller. 2014. Query-by-example image retrieval using visual dependency representations. In *COLING 2014*, pages 109–120.

Mark Everingham and John Winn. 2012. The pascal visual object classes challenge 2012 (voc2012) development kit.

Zoe Falomir, Ernesto Jiménez-Ruiz, M Teresa Escrig, and Lledó Museros. 2011. Describing images using qualitative models and description logics. *Spatial Cognition & Computation*, 11(1):45–74.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard H Hovy. 2013. Learning whom to trust with mace. In *HLT-NAACL*, pages 1120–1130.

Andrej Karpathy and Li Fei-Fei. 2014. Deep visual-semantic alignments for generating image descriptions. *arXiv preprint arXiv:1412.2306*.

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2011. Baby talk: Understanding and generating image descriptions. In *Proceedings of the 24th CVPR*. Citeseer.

Angeliki Lazaridou, Dat Tien Nguyen, Raffaella Bernardi, and Marco Baroni. 2015. Unveiling the dreams of word embeddings: Towards language-driven image generation. *arXiv preprint arXiv:1506.03500*.

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems*, pages 1682–1690.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Bernd Neumann and Ralf Möller. 2008. On scene interpretation with description logics. *Image and Vision Computing*, 26(1):82–101.

Benjamin Rosman and Subramanian Ramamoorthy. 2011. Learning spatial relationships between objects. *The International Journal of Robotics Research*, 30(11):1328–1342.

Kristoffer Sjöö, Alper Aydemir, and Patric Jensfelt. 2012. Topological spatial relations for active visual search. *Robotics and Autonomous Systems*, 60(9):1093–1107.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555*.

Ying-Hong Wang. 2003. Image indexing and similarity retrieval based on spatial relationship model. *Information Sciences*, 154(1):39–58.

Yezhou Yang, Ching Lik Teo, Hal Daumé III, and Yiannis Aloimonos. 2011. Corpus-guided sentence generation of natural images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 444–454. Association for Computational Linguistics.

Yuke Zhu, Alireza Fathi, and Li Fei-Fei. 2014. Reasoning about object affordances in a knowledge base representation. In *Computer Vision–ECCV 2014*, pages 408–424. Springer.