Exploiting Spatial Relational Knowledge for Visual Cognitive Tasks

Medhat M. Riad (medhat.riad@phil.uu.nl)
Utrecht University, Heidelberglaan 6
3584 CS Utrecht, The Netherlands

Florian Röhrbein (roehrbei@informatik.uni-bremen.de)
University of Bremen, Enrique-Schmidt-Str. 5
28359 Bremen, Germany

Nils Einecke (nils.einecke@honda-ri.de)
Julian Eggert (julian.eggert@honda-ri.de)
HONDA Research Institute Europe GmbH, Carl-Legien-Str. 30
63071 Offenbach am Main, Germany

Abstract

Reasoning about objects in a visual scene can be substantially improved if we have a representation that includes both information about the object properties and information about the relations that hold between them. For this, we have built a semantic-relational network that makes use of textual commonsense knowledge of both sorts and allows for inference within a Bayesian framework (Röhrbein, Eggert, & Körner, 2007). In order to enrich this graphical representation, we target at adding complementary information gained, for example, from sets of labeled images. We are especially interested in relational knowledge, a demand that rules out most existing image databases since they usually contain only one labeled object per image. LabelMe proved as a promising alternative and we started our efforts by analyzing the statistical dependencies between all objects in fully-labeled images of that database. In the contribution here, we report these results and discuss how the gained information can be used to build contexts as a first important step in reaching the final goal of learning a relational knowledge representation in an autonomous way.

Keywords: inter-object relations; spatial context; labeled images

Introduction

Visual scenes contain a wealth of information which would be very useful for technical systems. The complexity of visual information and the processing and storage costs involved in acquiring any form of this information that is directly usable, makes the use of visual information very challenging. This is further complicated by the problem of deciding what kind of information is useful and what is simply superfluous. However, a lot of research has been carried out over the years to obtain useful information from visual scenes whether it be high-level, global context information (Fei-Fei & Perona, 2005), object recognition (Wersing & Körner, 2003), or low-level image processing for feature detection (Lowe, 2004). In this paper, we report on preliminary results of statistical analysis on images at an intermediate-level where object information is assumed to be readily available showing that there are some statistical regularities in visual scenes at a level below that of global cues (like texture or gist models) and above that of local analysis (such as edge detection or single objects).

Working at object-level means that annotated images must be available. Some of the common analysis techniques include object positions, percentage area covered by object, and object complexity (boundary points). However, these techniques do not scale very well with number of images and if any regularities are seen then they are more representative of artistic preferences of photographers than actual object properties in real-world scenes (e.g. photographs of airplanes will usually be centered on the plane when obviously an airplane is seldom in the center of the visual field in everyday scenes). For this reason, it is our belief that analyzing object-object relations in images rather than object-scene relations would reduce the effect of photography bias and, as shown in the results, could detect some regularities in spatial relations between these objects.

The following section will give a brief introduction to the LabelMe database that we use in our analysis followed by an overview of region connection calculus, of which we used a modified version to describe relations between objects in images. In the results section, we report on some of the interesting results we found on different levels of detail. Following this, we dedicate the final section to discussing the results and how to interpret them as well as present the problems faced and how these problems affected the results. There is also a subsection discussing related work and the reported results and the discussion section ends by briefly talking about future plans and incorporating this information into our semantic-relational network.

Method

Annotated images

One of the main requirements for this research is to have a large database of images with multiple annotated objects per image. Even though there are several publicly available annotated image data-sets, most of them either have too few images for our purpose or contain only one object per image. Such single-object images are very useful as training data for object recognition but are useless for trying to uncover regularities in larger, complex, multi-object scenes. For this purpose, we use LabelMe (Russel, Torralba, Murphy, & Freeman, 2008) which is an on-line, publicly accessible database which depends on Internet users to upload and annotate images.

The LabelMe database consists of many images along with corresponding XML files containing the label names and polygon points representing each labeled object in the image. A single image may contain any number of labeled polygons (see Figure 1 for an example). Internet users can view current labels and add new ones. Labels of objects are typed in using the keyboard and there are no restrictions on what any label can be.

Undoubtedly, the use of such an "uncontrolled" data-set introduces some problems. One very pronounced problem is that there could be many variations of the same object class (e.g. *trashcan*, *bin*, *trash bin*, *waste basket*, etc...). In the reported results, all the different "synonyms" of an object class are treated as if they are independent classes without any attempt to normalize or consolidate the user-submitted labels. Another related problem is that of viewing-angle labels. Many users provide some hints in the labels as to what the viewing angle of the object is, and even though there are some attempts by the annotators to have a standard for providing such information, it is largely un-managed resulting in labels like *car front*, *car side view*, and *car az240deg* specifying the azimuth in degrees. Such labels are going to be very useful at a later stage when view-dependent statistics are required but again, at this stage, each different label is treated as an independent object class. There are also spelling mistakes (e.g. *wnidow*) and garbage labels found in the data but these do not appear often enough to pass as legitimate object classes according to our filtering criteria discussed next.

LabelMe is a very large database, and many of its images have few or no annotated objects. One way to deal with this problem is to only use fully-annotated images. An image is considered to be fully-annotated if at least 90 percent of its area is assigned to at least one annotation. Of course this is an arbitrary choice but it provides a large enough data-set while removing most images that do not have enough labeled objects. The resulting data-set of 4,955 images had an average of 16.7 objects per image ranging from 1 object to 266. To increase the reliability of the results, any object class that had less than 10 instances in the database, or only existed in one image was excluded from the analysis, in the latter case, regardless of how many instances were found.

Region Connection Calculus

In order to be able to extract useful relations between objects in images we need a standard way of describing the possible spatial arrangements between them. One way to do this would be to have directional relations like *left-of*, *above*, *behind*, etc, but this introduces the problem of granularity and fuzziness so a much simpler and unambiguous representation was needed.

Region connection calculus (RCC) is a qualitative spatial representation to abstractly describe regions and their relations to each other (Randell, Cui, & Cohn, 1992). One version of RCC, called RCC8, consists of 8 basic relations that are possible between two regions in two-dimensional space. The *disconnected* (DC) relation represents two separate objects; *externally connected* (EC) represents separate objects that are touching at an edge; the *equal* (EQ) relation means both objects are exactly equal in size and in the same position; *partially overlapping* (PO) means part of one object intersects with the other; *tangential proper part* (TPP) and *tangential proper part inverse* (TPPi) represent the cases where one object is completely contained within the other and is touching at some edge; and *non-tangential proper part* (NTPP) and *non-tangential proper part inverse* (NTPPi) also represent cases of one object contained within the other with the difference that no edges are touching. From these basic relations, different combinations can be built, resulting in new relations.

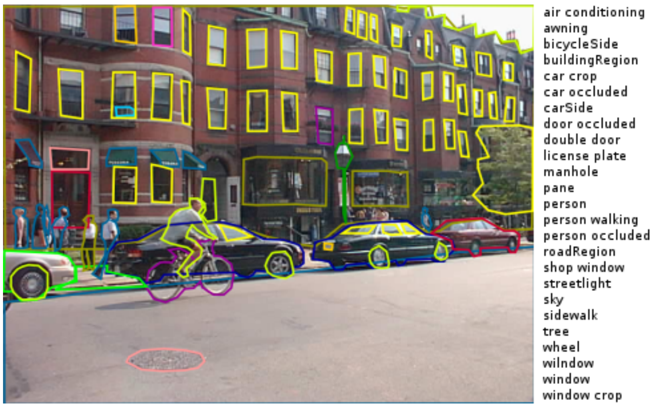


Figure 1: A fully-annotated LabelMe image with associated labels. In this example, *manhole* is a proper part of *roadRegion* and disconnected from *bicycleSide* which itself partially overlaps one of the *car occluded* objects (see next section). Repeated labels have been removed for brevity. (Image adapted from LabelMe (Russel et al., 2008))

In our analysis, we choose a subset of RCC8 which represents the combinations we found to be most likely in a two-dimensional projection of the three-dimensional physical world. The EC category is incorporated with the DC category by introducing a distance variable where a distance of zero represents EC and a distance greater than zero represents DC. We call this new relation, non-overlapping (NO). TPP and NTPP are combined into a proper part category (PP) which simply means that one object is completely contained within the other. Similarly, TPPi and NTPPi are combined into proper part inverse (PPi). The PO category is used as is with no changes while EQ is dropped all together since it would be highly improbable for two objects to be in exactly the same position with exactly the same size in an image and still be recognizable as to receive two separate labels.

Interestingly, Galleguillos, Rabinovich and Belongie (Galleguillos, Rabinovich, & Belongie, 2008) came up with similar spatial categories using a vector quantization approach. For the representation of spatial relationships they use INSIDE, AROUND, ABOVE and BELOW. The main difference to our RCC-derived categories lies in neglecting horizontal relations, which is due to the image databases they are using (see Discussion section).

For deciding which category an object pair falls into, polygon intersection areas are used. An intersection area of zero means an NO relation; PP or PPi is the result when the area of the intersection is equal to the area of one of the polygons but not the other, and a PO relation is the result when the intersection area is less than both polygon areas. For distance measurement, we use a minimum gap algorithm on the convex hulls of the corresponding object polygons and for angle measurement, we calculate the angle between the centroids of the polygons.

Results

In the following few subsections, results from different levels of abstraction are shown.

Object Co-occurrence

Using only fully-labeled images as described earlier, there was a total of 1,622 object classes which gives a total of 1,314,631 possible tuples. Out of these, only 3207 tuples were found to exist in the same image, which is less than 0.25 percent.

Conditional Probabilities

In an attempt to see how the presence of an object predicts the presence of another, we calculate the conditional probabilities of object 1 being found in an image given that object 2 is known to be there and vice-versa. It is not possible to list all these probabilities here so Table 1 shows all instances where this probability is one. In the table, the presence of “Object 1” in an image predicts with a probability of one, that “Object 2” is found somewhere in the same image. For example, the data predicts that if *taxi* or *van rear* is seen, then there is always a *road*. It has to be kept in mind that this does not make any predictions about how the objects are related spatially or conceptually, it is merely a co-existence correlation which explains why seemingly counter-intuitive pairs such as *one way sign* and *window* are so strongly correlated.

Table 1: All object pairs where the probability of appearance of the second object given the first object is present is 1

Object 1	Object 2	Object 1	Object 2
taxi	road	tray	wall
washbasin	wall	socket	wall
windshield (occ)	window	dishwasher	wall
van rear	road	cars side (occ)	road
wheel rim	wheel	telephone	wall
dishwasher	sink	ceiling	wall
light switch	wall	cushion	wall
dishwasher	faucet	bed crop	wall
cupboard	wall	toilet paper	wall
bed	wall	outlet	wall
electrical outlet	wall	one way sign	window
wheel (occ)	window	cabinet	wall
ceiling light	wall	motorcyclist	road
toilet	wall	rug	wall
alarm clock	wall	end table	wall
papers	wall	porch	window

Spatial Relations

Out of the 3,207 found co-occurring pairs, there are 144 pairs that have a probability less than 0.5 to be in the NO category. Only 10 pairs are found to belong to category PP with a probability greater than 0.5 (Table 2). The PPi category contains the same 10 pairs with a probability greater than 0.5 with the difference that “Object 1” and “Object 2” are switched. For the PO category, 103 pairs with probability greater than 0.5 are found.

Table 2: All pairs of the proper part (PP) category with probability greater than 0.5

Object 1	Object 2	Object 1	Object 2
bed	pillow	field	shrub
building region	window	house	sign
buildings	windows	road	manhole
car az30deg	wheel rim	road	manhole cover
ceiling	ceiling light	wall	alarm clock

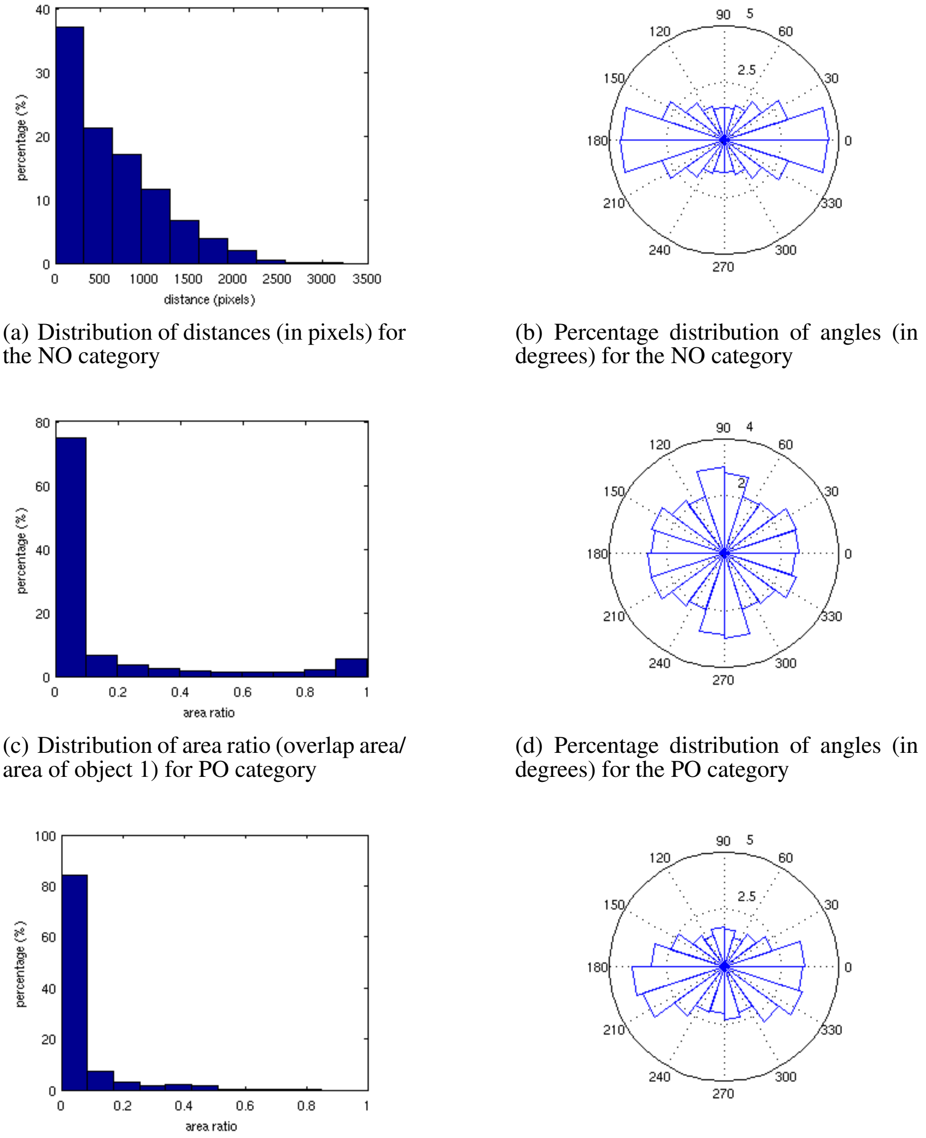
Distributions

For each of the categories, distributions of selected variables over the individual instances are presented here. For the NO category, distributions are shown for the distance[1] between objects as well as the distribution of the angle between them (Figures 2(a) and 2(b) respectively). For PO, Figure 2(c) shows the distribution of the ratio between the area of the overlapping regions and the area of the first object while Figure 2(d) shows the angles between the partially overlapping objects. For PP (Figures 2(e) and 2(f)), the distributions show the angles between the objects and how much of the larger object is covered by the smaller one (represented as a ratio). Distributions for PPi are the same as those for PP with the difference that angle distributions are mirrored along the horizontal direction.

Discussion

Some research suggests that many cognitive abilities, including abstract thought, make use of spatial relations or schemas in contexts that do not necessarily involve physical spatial properties (Gattis, 2001). An example would be people thinking about horizontal directions when thinking about *siblings* or the upwards direction when thinking about *respect*. It is therefore our belief that spatial relations are important in knowledge representations since they are not only confined

[1] Several image sizes have been used therefore the distance measurements also include an implicit depth attribute.



(a) Distribution of distances (in pixels) for the NO category

(b) Percentage distribution of angles (in degrees) for the NO category

(c) Distribution of area ratio (overlap area/ area of object 1) for PO category

(d) Percentage distribution of angles (in degrees) for the PO category

(e) Distribution of area ratio (area of object 2/ area of object 1) for PP category

(f) Percentage distribution of angles (in degrees) for the PP category

Figure 2: Various distributions between object pairs of the different relation categories

to relations between physical objects thus providing grounding for other, more abstract concepts.

In the physical context, it could at first seem that natural scenes are combinatorially infinite in terms of object relations, however, previous research (Torralba, Oliva, Castelhano, & Henderson, 2006) shows regularities in distributions for object instances of the same class. The results presented here, show that there are also regularities between object instances of different classes which limits the possible combinations, or at least, highlights the most probable ones. The results also show that more than 99.75% of all object pairs tested, never co-occur in the same scene, suggesting that it is not very likely that any of these pairs would occur in normal, everyday scenes. It is also predicted by the results that the majority of co-occuring objects in a scene will be nonoverlapping. Even though we believe that this prediction is overly exaggerated due to reasons discussed later, it seems that it is also a viable characteristic of natural scenes since most objects will, in fact, have some distance between them except in highly crowded scenes.

There are also some regularities seen regarding the relative positions and distances for the different relation categories. The distributions in Figure 2 seem to suggest that non-overlapping, co-occurring object pairs tend to be close together and positioned horizontally from each other. Partially overlapping pairs show a very small overlap ratio to the area of the first object and are more evenly positioned with a preference towards the vertical direction. For the proper part pairs, it seems that the majority of smaller objects are much smaller than the bigger ones. It might also be worth noting that the angle distributions for NO and PO are almost symmetrical in both the vertical and horizontal directions while the PP and PPi angle distributions are only symmetrical along the vertical. Even though it might not be straight forward to interpret what these regularities mean, it seems that these regularities do exist and can be helpful for visual tasks.

Looking at results from other research, it has been shown that multi-class object detection using contextual information between objects and object parts has shown an improvement in performance over systems which do not use such contextual information (Fink & Perona, 2004). However, this work has only been applied to face recognition tasks and it is not clear how this would generalize to other tasks. Also, the focus of this work is on the interactions between part and whole detection of one object rather than on the inter-object relations which our work focuses on.

There is also work on improving image segmentation by finding associations between image segments and a list of word tokens the image is annotated with (Carbonetto, Freitas, Freitas, & Barnard, 2004). The way this works is by formulating a probabilistic mapping between continuous image feature vectors and the supplied word tokens, which can lead to smoothed image segmentation, especially in situations where the scene was previously over-segmented. In contrast, we make use of already labeled multi-object images and directly analyze the spatial relations which hold between these objects within an image, giving inter-object relations rather than segment-word mappings.

A very recent paper by Galleguillos, Rabinovich and Belongie (Galleguillos et al., 2008) is most similar to our work reported here, since they make use of co-occurrence statistics as well as spatial relations to improve their object recognition system. They claim to be the first who make use of explicit spatial context between objects, but the generality of the results they obtain is severely limited by the databases (MSRC and PASCAL2007) they are using: MSRC comprises only 23 different categories while PASCAL2007 contains 20 categories. Even more severe, the images contain on average of only 3 (MSRC) or even 2 (PASCAL2007) objects. A further drawback with respect to accuracy is that both do not use polygons, but region masks or simple bounding boxes[2].

Problematic Cases

In our results, it is generally the case that many more instances of any given object pair fall into the NO category than we expected. For example, looking at *car* and *license plate* (see Table 3), we see that a *license plate* is only a proper part of a *car* about 25 percent of the time and non-overlapping with it the rest of the time with no partial overlaps. This might suggest that, 75% of cars do not have license plates which is obviously false. Such false predictions are made due to a combination of several factors, one of which is the aforementioned viewing angle. In this case, it could be said that out of all the times a car is seen, it is only seen at an angle which allows the license plate to also be seen, about 25% of the time. This makes sense if we assume that most of the time, cars are seen from the side. However, in other cases, viewing angle does not properly account for the discrepancy between the results and real world statistics. One of the most pronounced examples of this is that of comparing *building occluded* with *window* (see Table 3). In this example, we see a very high percentage of NO instances, with a relatively small percentage of PP instances where a window is completely inclosed within a building. One of the factors which brings about such inaccurate results is the inconsistency in the labeling. Some buildings will have every single window with a separate label, some buildings will have all of its windows labelled with a single all-inclusive label. Others still, will have only one or two labelled windows and sometimes none, and this obviously skews the results. Another very important issue is that of the category abstraction level of a label. In this case, the label “window” is found in the dataset to represent instances of car windows, building windows, indoor windows, and this contributes to a certain extent to the NO category overtaking the others. A more illustrative example might be that of the object *leg* which could represent a dog’s leg, a person’s leg, or a chair leg.

Some of the other problems we found include polygon inaccuracy and edge overlap issues. In polygon inaccuracy, labelers often need to estimate borders and there can always be a small overlap between separate objects or parts of a supposedly PP object lying outside of its enclosing object. In edge overlap situations, where one object completely hides the edge of another object, labelers tend to trace the visible border between the two objects rather than approximate the actual edge, giving inaccurate results. For example, in images with buildings that have cars parked in front of them, it is very common to see labelers trace around the cars instead of guess where the building would meet the ground.

There is also one particular problem that has proven to be not so trivial to deal with; and that is the problem of “false-positives”. Many pair instance comparisons result in an NO relation when in fact one of the instance pairs has a PO or a PP relation with another instance in the same image. For example, if we go back to our *car/license plate* example, having multiple instances of cars and license plates in an image is problematic, even if the labeling is accurate and complete. If we assume there are n cars and n license plates in the image and every instance of *car* has exactly one PP relation with a *license plate*, we can see that there are $n^2 - n$ NO relations and only n PP relations. And the problem gets worse with more instances since increasing the number to $n+1$ cars and license plates, gives a much higher increase in NO to $n^2 + n$ compared to the increase in PP which only becomes $n+1$.

[2] For a detailed comparison of the different labeled image databases see (Russel et al., 2008).

Table 3: Some inaccurate examples (Object names are followed by number of instances found)

	Example 1	Example 2
Object 1	car (101)	building (occ) (127)
Object 2	license plate (111)	window (603)
Tuples	151	833
NO	114 (≈ 75%)	709 (≈ 85%)
PO	0 (0%)	3 (≈ 0%)
PP	37 (≈ 25%)	121(≈ 15%)
PPi	0 (0%)	0 (0%)

In an attempt to solve this problem we separate the NO category into two sub-categories, NO-1 and NO-2, where NO-1 represents pairs where the first object does not overlap any instance of the second object in the same image and NO-2 represents pairs where the first object does not overlap a given instance of the second object while it does overlap at least one other instance. This distinction between "verified" NO pairs and "suspicious" ones gives the ability to have some idea of how strong the overall NO relation between the object classes is while not making any assumptions based on our knowledge of the objects in question. For the *license plate* example this gives 81 instances in the NO-1 category (≈ 54%) which is still quite high, but it must be noted that *license plate* can also be part of *van* or *car back* so some processing of the synonyms of object names could give much better results (e.g. using WordNet (Fellbaum, 1998)). It is still not very clear what NO-2 indicates since in some situations it would contain legitimate non-overlapping object pairs, and in other situations it would include the false-positives we wish to eliminate. We found no trivial, generic way to differentiate between the two situations without having some previous knowledge about the objects in question.

Future Directions

One of the reasons for performing this analysis is to extend a semantic-relational network with information about spatial relations between objects. It is relatively straightforward at this stage to use the results we have to store Bayesian spatial relations in the network. A future step would be to see how combining this setup with other techniques like saliency maps or feature detection algorithms would improve the performance in tasks such as object recognition or guided visual search. However, it must be noted that the results presented here are only for pairs of objects, and given the problems faced, there is still more work needed to improve the results (as discussed earlier). Another way the results could be made more reliable is by extending the idea to analyze triples or quadruples of objects. By moving towards incorporating more and more objects in the analysis, any regularities found would eventually form a context (for example *chair*, *desk*, *wall* and *window* can form a "working area" context which can be combined with other objects like *water dispenser* or *bed* to differentiate between offices and bedrooms). It seems plausible that the idea of a hierarchy of relations could prove to be very useful. Another way to improve the results would be to capitalize on orientation-dependent regularities that have already been found where it might be useful to use image analysis at an early stage to determine the pose of objects as it is very likely that pose-related regularities can be found (e.g. chair facing desk, mug handle facing outside). On a slightly different front, performing the analysis on sequences of images seems to be a very interesting idea, where the aim would be to discover regularities in the way spatial relations change over time and how these regularities might be different between different objects, adding an extra temporal dimension to the semantic-relational network.

References

Carbonetto, P., Freitas, N. de, Freitas, O. D., & Barnard, K. (2004). A statistical model for general contextual object recognition. In *ECCV* (pp. 350–362).

Fei-Fei, L., & Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. In *Proceedings of IEEE computer society conference on computer vision and pattern recognition CVPR'05* (Vol. 2, pp. 524–531).

Fellbaum, C. (1998). *Wordnet: An electornic lexical database*. Bradford Books.

Fink, M., & Perona, P. (2004). Mutual boosting for contextual inference. In *NIPS*. MIT Press.

Galleguillos, C., Rabinovich, A., & Belongie, S. (2008). Object categorization using co-occurrence, location and appearance. In *IEEE conference on computer vision and pattern recognition, CVPR-2008* (pp. 1–8).

Gattis, M. (Ed.). (2001). *Spatial schemas and abstract thought*. Cambridge, MA: MIT Press.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*, 91–110.

Randell, D., Cui, Z., & Cohn, A. (1992). A spatial logic based on regions and connection. In *Proceedings of KR'92* (pp. 165–176). San Mateo, California.

Röhrbein, F., Eggert, J., & Körner, E. (2007). A cortex inspired neural-symbolic network for knowledge representation. In *Proceedings of IJCAI'07 workshop on neural-symbolic learning and reasoning*. CEUR Workshop Proceedings.

Russel, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008, MAY). Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision*, *77*(1-3), 157–173.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological Review*, *113*(4), 766–786.

Wersing, H., & Körner, E. (2003). Learning optimized features for hierarchical models of invariant recognition. *Neural Computation*, *15*(7), 1559–1588.