Adding Spatial Semantics to Image Annotations*

Laura Hollink[1], Giang Nguyen[2], Guus Schreiber[1], Jan Wielemaker[2], Bob Wielinga[2], and Marcel Worring[2]

[1] Free University Amsterdam, Department of Computer Science
{hollink, schreiber}@cs.vu.nl
[2] University of Amsterdam, Informatics Institute
{giangnp, worring}@science.uva.nl, {jan, wielinga}@swi.psy.uva.nl

Abstract. In this paper we discuss a the support of users in adding spatial information semi-automatically to annotations of images. Descriptions of objects depicted in an image are extended with information about the position of those objects. We distinguish two types of spatial concepts: absolute positions of objects (e.g., east, west) and relative spatial relations between objects (e.g., left, above).
We show the use of a tool for a collection of art paintings with preexisting RDF annotations, including a list of image objects. First, the tool segments a painting into regions. The user selects regions, and labels these with objects from the existing annotation. Then, the tool computes absolute positions and relative spatial relations of the selected regions, and adds these to the annotation. A small evaluation study is reported in which annotations generated by the tool are compared to manual annotations by ten volunteers.

1 Introduction

In this paper we discuss semi-automatic annotation of images with spatial information. In a previous study [6] it was shown that people who describe images often use spatial descriptions like "On the left side" or "Below object x". Spatial information is important for describing the composition of an image, and for the identification of specific objects.

Making a complete and elaborate annotation of the content of an image is a time consuming process. Therefore, the human annotator should be supported in this task as much as possible. In spite of improvements in the field, automatic annotation of images is not feasible at the moment. This is due to the fact that what is depicted in an image is highly subjective. Spatial information, however, is mainly objective. This makes it a good starting point for semi-automatic annotation. This work can be seen as an exploration into bridging the "semantic

* An early version of this paper has been accepted for presentation at the 'Workshop on the Application of Language and Semantic Technologies to support Knowledge Management Processes' at EKAW 2004.

gap" [10], which refers to the cognitive distance between the analysis results delivered by state-of-the-art image-analysis tools and the concepts humans look for in images. In this work we use images from a collection of art paintings that we have used in an earlier study about semantic annotation [4]. The system we propose takes an annotated image as input. It segments the image into regions and allows the user to label the regions with concepts from the annotation. The system computes the position of the concepts and the spatial relations between them, and adds the spatial information to the annotation. A small evaluation is done in which annotations generated by our system are compared to manual annotations by humans.

It should be noted that this is an exploratory study to investigate the potential of content-based techniques for (spatial) image annotation at a conceptual level. As will be seen, we have deliberately "cut some corners" with the intention to show whether the idea could work in principle.

In the next section we discuss the representation of spatial information. In Sect. 3 we give a description of our system. Section 4 contains the results of a small evaluation study. The final section contains a general discussion.

2 Representing Spatial Relations

Talmy [12] describes spatial relations in the context of human perception. He conveys that the spatial disposition of an object in a scene is always characterized in terms of another object. The first object, which is called the 'figure', is the subject in the expression. The second object, or the 'ground', is used as a fixed reference to which the position of the figure is described. Grounds are for example the earth or the body of the speaker. More then one ground object is possible (e.g. "the bike is on the other side of the church": the bike is the figure, the church is the ground object, the body of the speaker is the second ground object). Another important point is that in human language a finite number of words is used to represent an infinite number of spatial configurations. This means that choices have to be made about which spatial concepts are used in a vocabulary.

Cohn [2] points out that when making a representation of space, questions have to be addressed regarding the kind of spatial entity being used (e.g. regions, points), and the way of describing relationships between these entities (e.g. their topology, size, distance, orientation or shape). For our practical purposes of annotating objects in images, we restricted ourselves to two-dimensional, binary relations between regions. The spatial relations that are included in our vocabulary must be (1) relevant for image annotations, and (2) suitable for automatic detection. This last requirement disqualifies concepts like 'behind' and 'in front of' since they are very hard to detect.

We distinguish two types of spatial concepts: absolute positions and relative spatial relations. The first are used to describe the position of objects within an image. The image functions here as the 'ground' of the expression. A common representation of absolute positions are the compass points North, South, East, West, Northeast, Southeast, Northwest and Southwest. We divided an image into nine squares where each of the outer squares represents one of the compass points and the middle square represents the center. Relative spatial relations are used to describe positions of objects relative to each other; one object is the 'figure', the other is the 'ground'. The set of relations that we used in this study includes: Right, Left; Above, Below; Near, Far; and Contains. One additional spatial relation can be derived, namely `Next` is either `Left` or `Right`.

In order to add the spatial information to semantic annotations of images, we used concepts from existing ontologies to specify the positions and spatial relations. Spatial relations were taken from SUMO [8]. This is a large, well structured ontology that takes into account Cohn's ideas about spatial relations.[1] Absolute positions were taken from the general lexical database WordNet [3]. One exception was the spatial relation `Far` that was taken from WordNet since it was not a concept in SUMO (version 1.15).

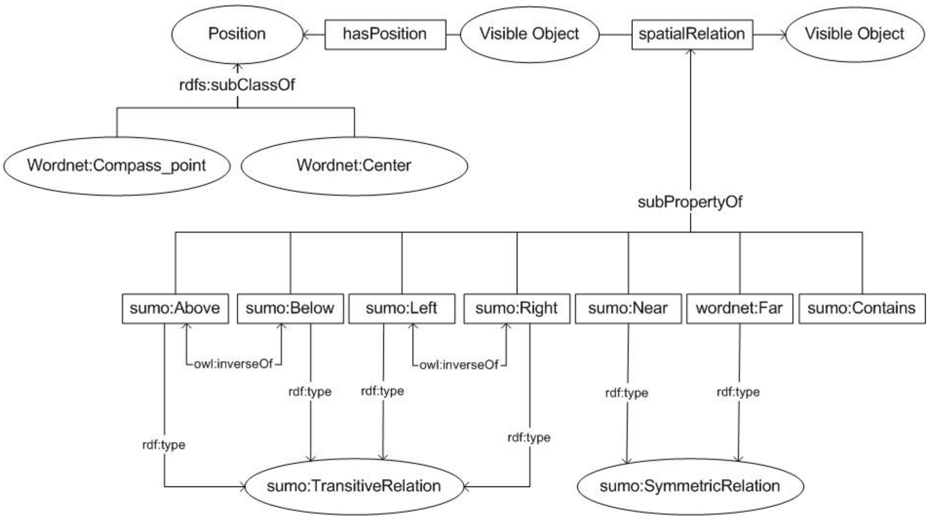


Fig. 1. Spatial concepts (ellipses) and their properties (rectangles) as they are used in our annotation schema.

For each spatial relation that we use we specify whether or not it is a `Symmetric Relation`, or a `Transitive Relation`, and what the `inverse Of` the relation is. RDF Schema is used for the representation of the spatial concepts[2]. Figure 1 depicts an RDF graph of the spatial annotation schema that we use. It shows a `Visible Object` that has a `Position`. The `Position` class has

[1] CVS log for SUO/Merge.txt, http://ontology.teknowledge.com/cgi-bin/cvsweb.cgi/SUO/Merge.txt, revision 1.24

[2] One term from OWL was used, owl:inverseOf, for there is no notion of opposite properties in RDF.

two subclasses, namely the WordNet classes `Compass Point` and `Center`. The `Visible Object` has a spatial relation with another `Visible Object`. We defined the spatial concepts from SUMO as subproperties of the property `spatial Relation`. `Left` and `Right` are each others inverse, just as `Above` and `Below`. All four are `Transitive Relations`. `Far` and `Near` are defined as being `Symmetric Relations`. We disregard Talmy here [12], who points out that near and far are in human language not used as symmetric relations: a bike can be near a house, but nobody will say that the house is near the bike. This has to do with the size and mobility of the objects, which are properties that we do not take into account at this time.

3 Spatial Annotation Tool

The system we propose helps the user to add spatial information to image annotations. For this purpose, we use a collection of art paintings that are annotated with the objects that are visible in them. The collection of images is first segmented off-line. For each painting color and texture features are extracted using Gabor filters. Pixels with similarity values above a given threshold are merged into a region. Several segmentations are computed for one painting, using different scales and thresholds.

The interactive annotation process consists of five steps: input, interactive segmentation, annotation, computation of spatial relations, and output. In the *input* step, the user selects a painting from the collection. In the *interactive segmentation* step the relevant objects in the image are identified. In this step we employ the framework described in Nguyen & Worring [7]. The system first offers the user a segmentation of the image using the default set of parameters. The user can now ask for a larger or smaller number of regions, after which the systems updates the parameters. This process goes on until the user is satisfied with the segmentation. By allowing the user to give feedback, the resulting segmented image will closely match the user's expectations. Different purposes require segmentations at different levels.

In the *annotation* step, meaning is added to the relevant objects. The user labels regions in the segmented image with concepts from the annotation. The labelling is done by clicking on a region and clicking on a concept from the annotation. Fig. 3 shows the interface of the system, at the moment that a user is labelling the regions. When the user decides that all relevant regions are labelled, the system continues to the *computation of spatial information* step. In this step, absolute positions and relative spatial relations of the selected regions are computed. Each selected region is represented by a bounding box and the center of the bounding box. Absolute positions are computed by determining in which of nine squares the center is. For the computation of the relative spatial relations we employ the method of Abella & Kender [1]. All relations are computed by comparing the centers and borders of bounding boxes of two objects. In Fig. 2 the definition of `Left` is shown as an example. For details of the other relations we refer to the reference.

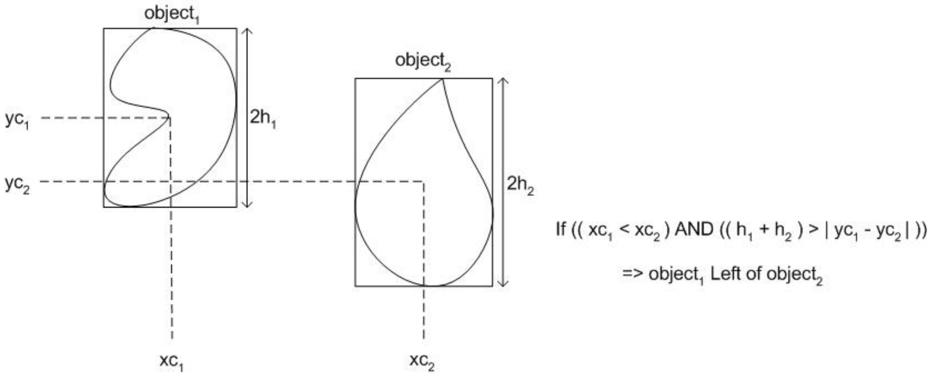

object_1
object_2
yc_1
yc_2
$2h_1$
$2h_2$
xc_1
xc_2
If (($xc_1 < xc_2$) AND (($h_1 + h_2$) > | $yc_1 - yc_2$ |))
=> object_1 Left of object_2

Fig. 2. Definition of the spatial concept `Left`.

Finally, in the *output* step, the spatial information is written as RDF statements to the original annotation, from where it can be queried by other tools. Fig. 5 depicts a screenshot of the Triple20 toolkit[3], that can be used to display and query the annotations. The figure shows the graphical output of Triple20 that displays the spatial annotation of the Matisse painting "Conversation" (Fig. 4) as an RDF graph. The annotation includes two objects linked by the SUMO concept `Left`. The position of one of the objects is specified by a WordNet concept with the meaning `East`.

4 Preliminary Evaluation

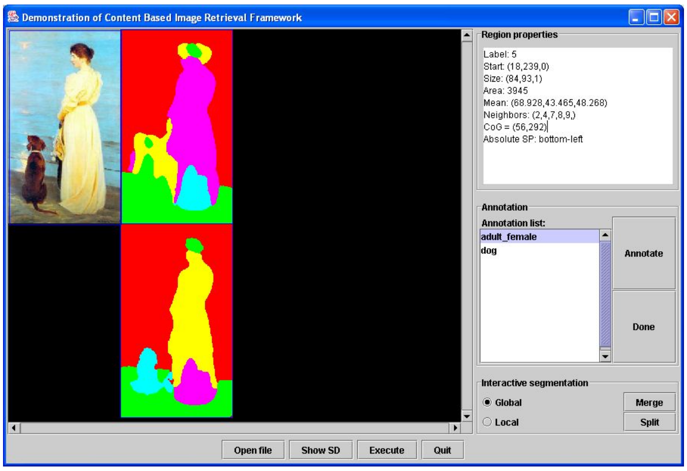
4.1 Methods

While designing the tool we have made decisions regarding the choice of concepts that are incorporated, and the definitions of these concepts. In this user study we evaluate these decisions. We asked two questions:

1. Are the spatial concepts that the tool uses the same as the concepts that users would use?
2. Are the definitions of the spatial concepts in accordance with the intuition of users?

Shariff & Egenhofer [9] asked similar questions for relations between lines and regions. They asked human subjects to draw sketches of English-language spatial terms. The sketches were used to map spatial terms onto geometric parameters and their values. One of their results was that topology was more important than metric properties in the selection of spatial terms. We took another approach:

[3] Triple20 is an open-source Prolog-based semantic-web package, see http://www.swi-prolog.org/packages/Triple20/.



Label: 5
Start: (18,239,0)
Size: (84,93,1)
Area: 3945
Mean: (68.928,43.465,48.268)
Neighbors: (2,4,7,8,9,)
CoG = (56,292)
Absolute SP: bottom-left

Annotation list:
adult_female
dog

● Global
○ Local

Fig. 3. Screenshot of the spatial annotation tool, showing a painting segmented at two levels. *Region properties* of the selected region are shown in the top right corner. Concepts from the annotation are listed in the *Annotation list*.

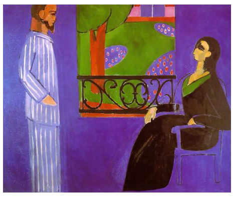


Fig. 4. "Conversation" by Henri Matisse, 1909

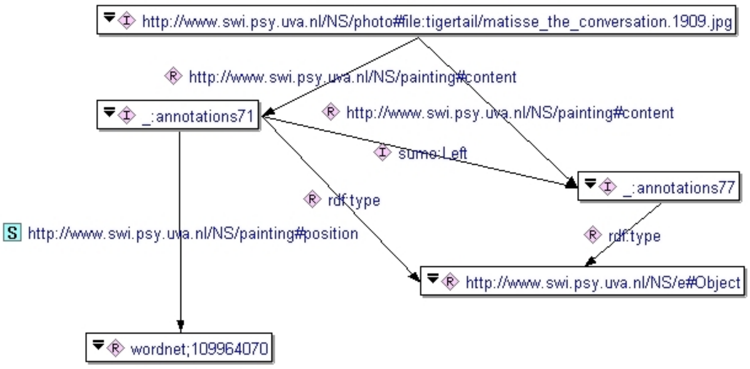

http://www.swi.psy.uva.nl/NS/photo#file:tigertail/matisse_the_conversation.1909.jpg
http://www.swi.psy.uva.nl/NS/painting#content
_:annotations71
http://www.swi.psy.uva.nl/NS/painting#content
sumo:Left
_:annotations77
rdf:type
http://www.swi.psy.uva.nl/NS/painting#position
rdf:type
http://www.swi.psy.uva.nl/NS/e#Object
wordnet;109964070

Fig. 5. Screenshot of Triple20's graphical output of a spatial annotation

subjects were asked to select spatial terms when provided with a configuration of objects in an image.

For the study we selected eight paintings that were well segmented by the tool (this seems a legitimate criterium since we are not evaluating the segmentation algorithms). Another criterium was that the paintings had to contain at least two objects. We asked ten PhD students who were familiar with annotation but not in particular with spatial concepts to participate in the study. They were split into two groups of five in order to answer the two evaluation questions.

Group 1 were provided with the eight paintings associated with a list of the objects that were visible on each painting. They were asked to provide statements about the absolute positions and relative spatial relations of these objects. Any number of statements was allowed. Comparing the spatial concepts that were used by Group 1 to the concepts included in the tool, will give an answer to Question 1.

Group 2 was also provided with the eight paintings and a list of objects. They were asked to describe positions and spatial relations using a limited list of spatial concepts. The list contained only the terms that are included in the tool. Again, any number of statements was allowed. Comparison of the statements of Group 2 to the statements of the tool will answer question two. We make the assumption that the spatial concepts that humans select are the correct ones.

4.2 Results

Group 1 In total, 257 statements were written down by Group 1: 129 absolute positions and 128 relative spatial relations (Table 1). 81 Percent of the absolute positions of Group 1 were concepts that were included in the tool. 8 Percent consisted of concepts that were not included in the tool. This were mainly three-dimensional positions such as "background" and "in front". The remaining 11 percent of the statements of Group 1 were more precise versions of the concepts in the tool. Examples are "almost in the center", "far right", "between left and center".

Of the relative spatial relations only 57 percent of the statements by Group 1 were concepts that were included in the tool. 29 Percent of the descriptions were concepts that were not in the tool; these were mainly three dimensional relations ("behind", "in front of"), statements about the connectedness of two objects ("connected", "freestanding") and "between". 14 Percent were more precise or less precise versions of concepts in the tool. "Object1 is northwest of Object2" is more precise than the concepts "above" and "left" in the tool, while "Object1 is higher than Object2" is more general than the concept "above" in the tool.

Group 2 The five subjects of Group 2 produced a total of 234 statements. Together they selected 127 absolute positions of 27 objects (Table 2). Of the 127 positions, 88 (69 %) matched the absolute positions that the tool computed. 39 Positions did not correspond to the computed positions, which seems a high number of mistakes. However, note that the tool cannot match all statements

Table 1. Summary of the results for Group 1, divided over absolute positions (AP) and relative spatial relations (SR)

Group 1	**AP**	**SR**	**Total**
Included in the tool	107 (81 %)	70 (57 %)	177 (69 %)
Not included in the tool	11 (8 %)	36 (29 %)	47 (18 %)
Not precise enough in the tool	14 (11 %)	18 (14 %)	32 (13 %)
Total	132 (100 %)	124 (100 %)	256 (100 %)

when the participants disagree about the position of an object. We found that for only seven of the 27 objects a majority of the participants (at least 3) agreed on a position different from the tool's position. An example of such a mistake by the tool is the window in the Matisse painting *Conversation*. The tool assigned the window the position *North*, while all subjects agreed that it was in the center.

Group 2 produced 107 statements about relative spatial relations. Not all possible relations between two objects were described by the subjects. It appeared that they used the `inverse Of` and `symmetric Relation` properties for the selection of relevant object pairs: when a subject had stated "woman left of man", he or she would not also state "man right of woman". To make the statements comparable to the statements of the tool, that did compute relations between each object pair, we added symmetric and inverse relations where necessary. This brought the total number of relative statements of Group 2 to 210 (and the total number of statements of Group 2 to 337). 154 Of these (73 %) were also found by the tool, 56 (27 %) were not.

Table 2. Summary of the results for Group 2, divided over absolute positions (AP) and relative spatial relations (SR)

Group 2	**AP.**	**SR.**	**Total**
Found by the tool	88 (69 %)	154 (73 %)	242 (72 %)
Not found by the tool	39 (31 %)	56 (27 %)	95 (28%)
Total	127 (100 %)	210 (100 %)	337 (100 %)

Another evaluation measure is the proportion of statements of the tool that corresponds to statements of the subjects. The tool computed 106 statements. 24 Of these were about an object pair that was not described by any of the participants, which means they cannot be validated. Of the remaining 82 statements, 56 (68 %) corresponded to at least one participant. Of the 26 'incorrect' statements of the tool, 18 concerned `far` and `near`. Participants hardly used these concepts.

5 Discussion

In this paper we explored the possibility to use a content-based image analysis technique to aid the process of spatial image annotation. The study shows there are indeed some points where the "semantic gap" can be bridged. A number of spatial concepts specified by human annotators were compatible with annotations produced by the tool. The results of the study seem to indicate that the absolute positions in the tool are roughly the same as the concepts that human annotators use. However, a number of relative spatial relations that people tend to use are missing from the tool. The choice of the set of spatial concepts was based on pragmatics, namely those for which automatic detection methods were available. The evaluation showed that this is a severe limitation since people often use three dimensional concepts, which are very hard to detect. Other frequently used concepts that the tool could not handle were `connected` and `between`. We are planning to include those in the next version of the spatial annotation tool. Two concepts included in the tool were hardly used by human annotators: `Far` and `Near`. It would be interesting to see whether this is also the case in other domains than art paintings.

The tool detected almost three quarters of the spatial concepts selected by humans. The results for relative spatial relations were slightly better than for absolute positions. This could be due to the fact that the tool assigns one position to each object, while any number of spatial relations can be detected for one pair of objects. This makes it possible to match all statements, even if subjects disagree with each other.

This was just an exploratory study with the aim to see whether this approach could work in principle. We can see the following lines of research as interesting follow-up options. Firstly, one could think of extending the functionality of the image-analysis tool to include a larger set of spatial relations. In the short term, this is likely to be limited to two-dimensional relations. Secondly, we should include ontological reasoning to derive spatial relations from the existing annotations. Such functionality is currently not included. Thirdly, one could consider including facilities for manual segmentation. This could improve the quality for images that are segmented badly by automatic techniques. Ley [5], for example, uses SVG to manually define regions and then annotates each region. Finally, it would be worthwhile to consider whether the content-based segmentation can also be used for other annotation purposes. One can think of other non-spatial properties of which the value can be derived with the help of segmentation. One example would be the color of a particular object. In the VisualSEEk system [11], for example, query by sketch is done based on colors and (relative) spatial locations of regions in an image.

Acknowledgements

This work is supported by the project "Interactive Disclosure of Multimedia Information and Knowledge" funded by the IOP Programme of the Dutch Ministry of Economic Affairs. We thank the subjects for investing their precious time in our evaluation study.

References

1. A.Abella and J.R.Kender. From images to sentences via spatial relations. In *Proc. of the ICCV'99 Workshop on Integration of Image and Speech Understanding.*
2. A.G. Cohn and S.M. Hazarika. Qualitative spatial representation and reasoning: An overview. *Fundamentae Informaticae*, (46):2–32, 2001.
3. C. Fellbaum, editor. *WordNet: An Electronic Lexical Database.* Bradford Books, 1998.
4. L. Hollink, A.Th. Schreiber, J. Wielemaker, and B. Wielinga. Semantic annotation of image collections. In *Proc. of the K-CAP 2003 Semannot Workshop*, Florida, USA, October 2003.
5. J. Ley. Raster image description and search in svg. Presented at the third annual conference on Scalable Vector Graphics (SVG Open): http://www.jibbering.com/svg/talk2004/title.html, Tokyo, Japan, September 2004.
6. L.Hollink, A.Th.Schreiber, B.Wielinga, and M.Worring. Classification of user image descriptions. *Int. Journal of Human Computer Studies*, November 2004.
7. G.P. Nguyen and M. Worring. Query definition using interactive saliency. In *Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval*, Berkeley, CA, USA, 2003.
8. I. Niles and A. Pease. Towards a standard upper ontology. In Chris Welty and Barry Smith, editors, *Proc. of FOIS-2001*, Ogunquit, Maine, October 17-19.
9. A. Rashid, B.M. Shariff, and M.J. Egenhofer. Natural-language spatial relations between linear and areal objects: The topology and metric of english-language terms. *Int. Journal of Geographic Information Science*, 12(3):215–246, 1998.
10. A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12), December 2000.
11. J.R. Smith and S-F. Chang. Visualseek: a fully automated content-based image query system. In *Proceedings of ACM Multimedia*, pages 87–98, Boston, MA, November 1996. ACM Press.
12. L. Talmy. How language structures space. In H. Pick and L. Acredols, editors, *Spatial Orientation: Theory, Research and Application*, New York, 1983. Plenum Press.