In *Working Notes of the AAAI Workshop*
on Integration of Natural Language and Vision Processing
AAAI-94, Seattle, Washington, 1994.

Integrating Text and Graphical Input to a Knowledge Base *

Raman Rajagopalan
Department of Computer Sciences
University of Texas at Austin
Austin, Texas, 78712

Abstract

The saying 'a picture is worth a thousand words' exemplifies the great value of pictures in describing a scenario. Pictures convey spatial information in a compact form, allowing textual descriptions to concentrate on the contextual properties of objects. The difficult task in integrating text and diagrammatic input to a system is to establish *coreference* - matching object references in the text to objects in the diagram.

We describe a tool, the Figure Understander, which has been developed for integrating text and graphically-drawn diagrammatic input to a problem solving system. Graphical editors not only provide a geometric description which easily identifies the objects in a scene, but also provide other features, such as colors and filling patterns, for distinguishing objects. We use these features to provide limited semantic information about objects directly in the diagram. We show that this greatly simplifies the solution to the coreference problem.

1 Introduction

Figures, consisting of a picture and an accompanying text description, are heavily used in both formal and informal modes of human communication. Pictures are used to illustrate objects and their spatial attributes and relationships, and text is used to describe the conceptual properties of those objects. The unique roles of diagrams and text has already been analyzed in great detail by such authors as Larkin and Simon [1987]. To integrate the information in text and diagrams, one must solve the problem of establishing *coreference* [Novak and Bulko 90, 93] - matching object references in the text against objects in the diagram.

We are developing a problem solver for the magnetic fields domain which uses qualitative spatial and dynamic reasoning to solve the problems [Rajagopalan 94]. One of our requirements was to develop a method for providing spatial, semantic, and dynamic change information about objects as input. We found that special purpose text forms were too cumbersome for providing a complete description of the spatial state, and that diagrams were a far more natural method of providing the same information.

Since the primitive objects in a diagram are lines, rectangles, ellipses, and polygons, we needed another form of input to describe the purposes of the objects and any changes that may be taking place. In written human communication, text is used to provide the missing conceptual information. Textual references to objects, however, are given in terms of semantically meaningful labels, such as magnetic field, making it difficult to locate the corresponding diagram object. Although there may be some clues in the text, such as references to spatial relationships between objects, to assist in matching diagram and text objects, it is often assumed that the reader has the necessary 'commonsense' understanding of the domain of discourse to make the connection between textual references and diagram objects.

Consider the scene in Figure 1. Here, the picture shows a circular object, a vector, and two rectangular objects. The textbook description [Resnick and Halliday 88, Chapter 32, Question 7] of the associated magnetic fields problem reads:

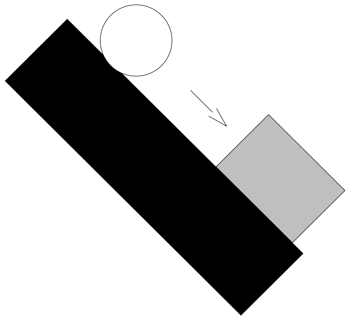
> Figure X shows an inclined wooden track that passes, for part of its length, through a strong magnetic field. You roll a copper disk down the track. Describe the motion of the disk as it rolls from the top of the track to the bottom.

Although this verbose description provides a lot of information, it does not directly identify the circular object as the rolling disk, or specify that 'copper' objects are conducting objects and that 'wooden' objects are non-conducting objects. These associations are trivial for the human reader, but can be difficult for an automated text and diagram understanding system.

Our goal was not to solve difficult problems in either visual scene interpretation or natural language processing, but rather to build a tool to make input descriptions easier for our problem solving system. Once we decided to input the spatial state through a diagram, we found

*This work has taken place in the Qualitative Reasoning Group at the Artificial Intelligence Laboratory, The University of Texas at Austin. Research of the Qualitative Reasoning Group is supported in part by NSF grants IRI-8904454, IRI-9017047, and IRI-9216584, and by NASA contracts NCC 2-760 and NAG 9-665.

- **Diagram:**



- **Text:** The disk is rolling (down the track into a magnetic field).

Figure 1: Example of text and diagram input for the magnetic fields problem solver. The diagram describes the spatial state and the text describes the dynamic changes that are taking place. The parenthesized words are unnecessary since we specify semantic information directly in the diagram.

that we also had to address the research issues associated with integrating multi-modal input [Novak and Bulko 90, Srihari 91]. Whether we chose to provide the information missing in the diagram through text or direct knowledge base assertions, we still needed a means of associating object references in the second form of input with the appropriate diagram objects. We found that the problem descriptions in our application could be reduced to a few simple sentence types and opted to use natural language text as the second input medium.

1.1 The Figure Understander

In this paper, we describe the Figure Understander, an implemented research tool for integrating the information in text and *graphically drawn* input into a single knowledge base description. The Figure Understander takes as input a diagram in the form of a PostScript file produced by an object-oriented graphical editor, such as the InterViews drawing editor, and one or more sentences of descriptive text. Its output is a fragment of a knowledge base which contains an integrated description of the information in each form of input. In the current implementation, the Algernon knowledge representation language [Crawford and Kuipers 91] is used to maintain the knowledge base. The Figure Understander makes use of a spatial reasoning module, a domain model, and a picture description model, as described further below, to process the text and diagram input.

The previous work in the area of integrating text and pictoral information [Novak and Bulko 90, Srihari 91] has assumed that the pictures were provided in visual (raster) form. Under such circumstances, visual scene analysis [Ballard and Brown 82] is required to identify the diagram objects of interest, and domain libraries covering the expected contents of diagrams have to be employed to determine the conceptual purpose of the diagram objects.

In contrast, graphical editors output geometric descriptions which directly identify diagram objects, and provide features (colors, filling patterns, brush types) other than just shapes and positions to distinguish objects. Our solution makes use of these additional features to provide limited semantic information for objects directly in the diagram. This simplifies the solution to the coreference problem by providing a common means of comparison - the conceptual class of objects.

For example, in Figure 1, we use a picture description language in conjunction with a magnetic fields domain model to define that 'black' objects are immobile, nonconducting, supporting blocks, that 'white' objects are mobile conducting disks, that 'gray' objects are magnetic fields, and that 'solid, thin vectors' represent directions of motion. By providing this semantic information directly in the diagram, the only additional initial state information required is that the disk is rolling, as shown in Figure 1. We omit the problem statement since it is always to describe 'what happens?'.

The Figure Understander first processes the diagram and outputs the assertion that the circular, white object is a conducting disk. Then, when processing the text, the reference to 'disk' generates a query to locate a conducting disk in the knowledge base. The reference to 'rolling' generates a query to find a vector which represents a direction of motion. Once the knowledge base objects are located, the Figure Understander asserts the new information that the disk is moving, and that its direction of motion is that given by the vector.

We have previously described the application of the Figure Understander as the input processing module for our magnetic fields problem solver [Rajagopalan and Kuipers 94]. In that application, the number of objects in the scene was small, each diagram object represented a conceptual object, and there was only one diagram object of each conceptual class in a given problem. Once the Figure Understander determined the conceptual class for a text object (through semantic interpretation functions), the solution to the coreference problem became straight-forward. The appropriate diagram objects could be accessed based on their conceptual type, and any additional information given in the text could be added to their knowledge base descriptions.

In this paper, we consider more complex examples where there can be more than one object of a given conceptual type in the diagram. These cases require additional information in the text, such as spatial relations, to identify the diagram objects of interest.

We describe the spatial representation method developed for our qualitative spatial reasoning (problem

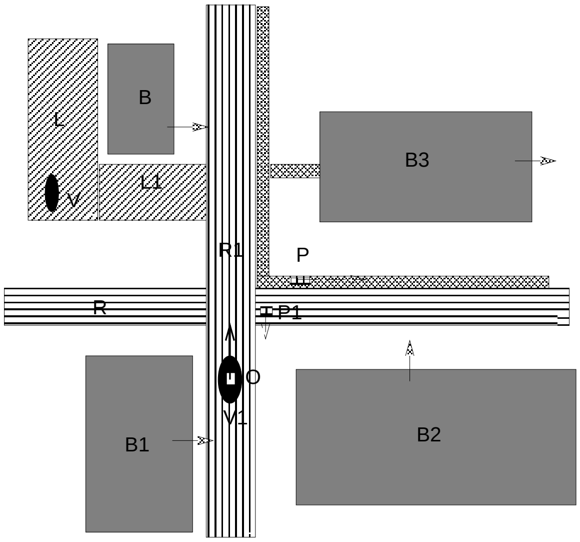


Figure 2: An representative input diagram: An urban scene - buildings, streets, vehicles, and pedestrians. The labels identify the internal names of the knowledge base objects created for each diagram object, and are strictly for the use of the reader. Figure Understander treats all text in diagrams as comments.

solving) application and explore its usefulness for natural language processing and diagrammatic reasoning applications - specifically, in the context of a urban scene understanding example. In the remaining sections, we first discuss issues related to diagram processing (Section 2), then discuss the spatial representation (Section 3) and its use during text processing (Section 4), and conclude with an example which illustrates the combined use of spatial relations in the text and semantic information in a diagram to solve the coreference problem (Section 5).

2 Diagram Processing

Figure 2 shows an input diagram that illustrates an urban scene consisting of buildings, streets, vehicles, and pedestrians. Notice that the diagram is only a collection of patterned or shaded vectors, rectangles, and ellipses. The fact that the objects are buildings, vehicles, etc., is revealed only in the accompanying text, as shown in Figure 3. Additional domain specific terms included in the text, such as Raman, Ben, Taylor Hall and Speedway Street, are labels which need to be attached to specific diagram objects.

The first step in our methodology for integrating diagram and text input is to associate a semantics with the objects in the diagram. We accomplish this by providing a method for the application builder to map the use of specific brush types, shading, and patterns in the diagram to specific domain object types.

2.1 The Picture Semantics Description

The `picture semantics` form, illustrated in Figure 4, is used to specify the connection between the patterns, colors, and brush types used in the diagram and the domain semantics.

The keywords, i.e., categories, `:patterns-for-2-d-objects` and `:brush-types-for-vector-objects` allow the user to specify an object type label for each of the color, pattern, and vector types allowed by the graphical editor used to draw the diagram. We use the `InterViews` graphical editor. In Figure 4, we include the information that a solid-white pattern represents an observer (speaker or listener), that a solid-gray pattern represents a building, and that a very light, solid-line vector drawn with the head pointed outward from the origin of the vector represents the intrinsic front direction of an object.

The keyword `:object-groups` is used to specify object classes from the individual types given in `:patterns-for-2-d-objects` and `:brush-types-for-vector-objects`. In Figure 4, we have defined the class 'roads' and have established that EW-road and NS-road are members of this set. In addition, we also specify the internal names to be used for objects of the class. In Figure 4, we have specified that all roads should be named 'R'. Thus, the first road identified in the diagram will be named 'R', the second 'R1', the third 'R2', and so on.

The object semantics are specified in two ways. Properties common to an object class as a whole are given through `:object-group-semantics`. The properties of individual object types within a class are given through `:individual-object-semantics`. In Figure 4, all semantic information is given in terms of Algernon relations. Under `:object-group-semantics`, we have specified that all diagram objects of the type 'roads' are members of the knowledge base class 'roads', and under `:individual-object-semantics`, have specified that the direction-of property of diagram objects of type EW-road is EW.

Under `:special-procedures`, application-specific diagram processing functions may be specified. These are run immediately after Figure Understander has completed its diagram processing task. In this example, the special procedure 'connect-vectors-to-objs' attaches the directional information given by the vectors in the diagram, such as the intrinsic front direction, to the objects which contain the tail of the vector.

3 The Spatial Representation

As the diagram is processed, a knowledge base (KB) object is created for each diagram object. Both the semantic information given in the picture semantics and the qualitative spatial properties of interest, such as

1. My name is Raman.
2. The name of the building that is to my right is Welch Hall.
3. The name of the building that is in front of Welch Hall is Taylor Hall.
4. The name of the building that is behind Taylor Hall is the Aerospace-Engineering building.
5. The name of the building that is the right of the Aerospace-Engineering building is the Biology building.
6. I am traveling on Twenty-Fourth street.
7. Speedway street intersects Twenty-Fourth street.
8. Ben is crossing Speedway street.

Figure 3: A collection of descriptive sentences for the scene depicted in Figure 2.

connectivity, orientation, and relative position, are attached to the KB description of the object. Human errors in drawing the diagram are taken into account by considering two coordinate values within a specified epsilon measure to be equal, and by considering two lines with nearly equal slopes to be parallel.

The Figure Understander extracts and reasons with a *qualitative* description of the initial spatial state. The goal is that any conclusions based on this qualitative description should apply to any physical initial state that fits the description, and not just the specific spatial state given by the diagram. We found that, in general, it is not possible to describe the exact shapes of objects in a qualitative manner, and like much of the previous work in qualitative spatial reasoning, use simplifying abstractions to approximate the spatial extent of objects.

Our magnetic fields problems involve reasoning about changes in the relative positions and orientation of objects due to either translational or rotation motion. Therefore, it was imperative that the chosen abstraction method could be described and *maintained* in a qualitative manner.

Previously developed spatial abstraction methods, such as the convex hull [Cui, Cohn, and Randell 92], inertia tensor-based bounding rectangles [Abella and Kender 93], bounding rectangles based on intrinsic front directions [Retz-Schmidt 88 (CityTour), Mukerjee and Joe 90], the idealizations based on outlines and volumes suggested by Herskovits [1985], or the cuboid shapes used by Maass [1994], are all models which can be easily computed given a *quantitative* description of the initial spatial state. However, none of the authors provide a *qualitative* means to describe the abstraction itself, particularly if quantitative knowledge of the initial state is unavailable, or describe how to maintain the position or orientation of the abstraction and the object within if the underlying object is free to translate and rotate.

Our spatial representation is based on knowledge of only the six *extremal points* of the boundary of an object: the rightmost, leftmost, topmost, bottommost, frontmost, and rearmost points. For problems involving static worlds or only translational motion, we use a rectangular bounding box or cube drawn around its extremal points to approximate the actual spatial extent of an object. We use bounding circles and spheres for problems involving the rotation of an object about an axis passing through its center of gravity. We provide a detailed discussion of the advantages and limitations of our spatial abstraction methods in [Rajagopalan 94].

We define the `orientation` of a surface through its surface normal direction. We infer the orientation of a surface through knowledge of the *identities* of the extremal points between which a surface lies. For example, the orientation vector of any surface between the topmost and rightmost points of an object will have positive X and Y components.

We model both directional (e.g., right-of, left-of) and topological (e.g., next-to, adjacent, overlapped) spatial relations between objects in terms of *inequality relations* between their extremal points. Our definitions represent only 'ideal meanings', as used by Herskovits [1985]. For example, we model that some object, A, is to the left of another object, B, through the inequality relation (rightmost(A) $<$ leftmost(B)).

Thus, not only can we describe the spatial state in a purely qualitative fashion, but our representation also allows us to easily recognize possible changes in the spatial state during translational and rotational motion. In the previous example, given that object A is moving to the right, we can infer that a qualitatively significant state change will occur when (rightmost(A) = leftmost(B)). We have also developed methods for inferring changes in the orientations of objects as they rotate.

3.1 Intrinsic/Deictic Frames of Reference

Retz-Schmidt [1988] discusses that in human use of spatial prepositions, extrinsic, intrinsic, or deictic

```
(defpict roads-and-buildings
    :patterns-for-2d-objects  ;Graphical Pattern  Domain Object Type
                             ((solid-white         observer)
                              (solid-black         vehicle)
                              (solid-gray          building)
                              (slash               parking-lot)
                              (horizontal-line     NS-road)
                              (vertical-line       EW-road)
                              (wide-rect-mesh      pedestrian)
                              (light-cross-mesh    sidewalk))

    :brush-types-for-vector-objects  ;Graphical Brush Type, Domain Object Type
                                    ((vl-s-vector-out      int-front-vect))

    :object-groups  ; Domain-Class-Name    Abb-Name   Member-Object-Type
                   ((observers               O         (observer))
                    (pedestrians             P         (pedestrian))
                    (roads                   R         (EW-road NS-road))
                    (buildings               B         (building))
                    (int-front-vectors       IF-V      (int-front-vect)))

    :object-group-semantics     ; Semantics given as Algernon predicates.
                            ((buildings    ((member buildings ?obj)))
                             (roads        ((member roads ?obj)))
                             (pedestrians  ((member people ?obj)))
                             (observers    ((member observers ?obj))))

    ; format is (class-name (object-type semantics) (object-type semantics) ..)
    :individual-object-semantics
                             ((roads
                               (EW-road ((direction-of ?obj EW)))
                               (NS-road ((direction-of ?obj NS)))))

    :special-procedures ((connect-vectors-to-objs))  )
```

Figure 4: Example of a partial picture semantics for urban scene interpretation.

frames of reference could be used. The Figure Understander represents spatial relations with respect to the extrinsic frame of reference of the diagram. The example sentences in Figure 3, however, require reasoning with intrinsic (sentence 3) and deictic (sentence 2) frames of reference. To solve this problem, the Figure Understander's spatial reasoning module allows queries about directional spatial relations to carry a frame of reference argument, and automatically determines an equivalent relation to prove in terms of the extrinsic frame of reference of the diagram. Currently, we only consider intrinsic and deictic frames of reference that are orthogonal to that of the diagram. Examples are provided in the next section.

4 Text Processing

Even with the ability to specify semantic information about objects in the diagram through the use of patterns and shading, it may not be possible to capture all domain information within the diagram itself. As discussed in Section 1, text may also be required. The example of Figures 2 and 3 illustrates the use of Figure Understander in associating labels given through text with objects in the diagram.

The task of attaching information given in an accompanying text sentence to diagram objects requires one to first locate the appropriate diagram objects. Since the picture semantics provides the identity of the conceptual class for each diagram object, we obtain the conceptual class of each text object through semantic interpretation functions and use that to match text objects against diagram objects.

In cases where there is only one diagram object of a given type (e.g., in Figure 2, there is only one speaker, O), the matching process becomes straightforward. If there are multiple objects of the same domain class in the diagram, we expect the text to provide additional information to narrow the choices to the appropriate diagram object. One common source of such information

```
(interpret '(the name of the building that
             is to my right is Welch Hall))

 The following Algernon path shall be asserted:

((:the ?V1 (MEMBER OBSERVERS ?V1))
 (:the ?V2 (MEMBER BUILDINGS ?V2)
           (:bind ?VECT
                  (GET-INT-FRONT-DIR-IF-ANY '?V1))
           (DIR-REL-WITH-FOR STRICTLY-RIGHT-OF ?V2
                             ?V1 SCENE-INIT ?VECT))
 (NAME-OF ?V2 '(WELCH HALL)))
```

Figure 5: Example of an input text sentence and the Algernon query and assert form generated by the text interpreter.

is the use of spatial relations in the text.

For example, consider the diagram in Figure 2 and the text sentences given in Figure 3. If we wish to interpret the sentence "The name of the building that is to my right is Welch Hall", we can use the fact that the word 'my' refers to the speaker, locate the only speaker (O) in the diagram, and then use the spatial relation 'building strictly to the right of the speaker' to locate the only building (B2) that satisfies the relation and label that as 'Welch Hall'.

4.1 Text Processing - Implementation

The text processing is accomplished by parsing the sentence through an ATN grammar, and then interpreting the parse output through a semantic interpreter. The semantic interpreter returns a query-and-assert form, currently in the syntax of the Algernon language, for locating the appropriate diagram objects in the knowledge base (query) and attaching any new information given in the text about the object (assert).

Figure 5 shows the Algernon form generated after processing the sentence 'The name of the building that is to my right is Welch Hall'. This sentence was interpreted by a semantic interpretation function which processes sentences of the form: (<property> of <subject> that <spatial relation> <reference object> is <value>).

There are two queries and an assertion. The first form, `(:the ?v1 (member observers ?v1))` is a query to locate a *unique* diagram object whose conceptual type is 'Observers' (speaker). The second (long) form is a query to find a unique building that is strictly right of the observer (based on intrinsic front direction of the observer). In Figure 2, the intrinsic front directions for the buildings and the speaker are given through the attached vectors. The final form in Figure 5 specifies a new fact to be added, the name of the building, should the query identify the appropriate building. In this case, building B2 will be named as Welch Hall.

The form `(GET-INTrinsic-FRONT-DIRection-IF--ANY)` returns the intrinsic front direction of an object if one exists. The form `(prove-DIRectional--RELation-WITH-given-FrameOfReference)` accepts a directional spatial relation, the identities of two objects, a scope argument (in Figure 5, the initial spatial state), and a frame of reference argument and proves a corresponding relation with respect to the extrinsic frame of reference of the diagram.

In the example of Figure 5, the front direction of the speaker is that of the diagram, and the Figure Understander directly proves the strictly-right-of relation between the speaker, O, and the building B2. However, when processing sentence 4 in Figure 3, since the intrinsic front vector of Taylor Hall (building B3 in Figure 2) points to the right, the query of the directional relation 'strictly-behind B3' translates to the query 'strictly-left-of B3' in terms of the extrinsic frame of reference of the diagram. Building B is then identified as the Aerospace Engineering building.

5 An Annotated Example

The previous sections have described the rationale and theory behind the implemented code. The following illustrates the processing of the diagram and text given Figures 2 and 3. In the presentation below, all computer generated output is shown in typewriter font.

```
> (parse-figure
    "speedway-24th.ps"       ; input diagram
   '((my name is Raman))     ; input text
   'roads-and-buildings))    ; defpict form

************************************************
***             Parsing The Figure           ***
************************************************

Extracting spatial information from the figure.
Executing user defined special procedures.
Interpreting text sentence:
 (MY NAME IS RAMAN)

 The following Algernon path shall be asserted:
((:THE ?V1 (MEMBER OBSERVERS ?V1))
 (NAME-OF ?V1 'RAMAN))
```

Figure Understander is invoked through the `parse--figure` function. The diagram is provided in the form of a PostScript file produced by the InterViews drawing editor. The input also includes a list of the sentences to be processed, and the picture semantics form to the used. As Figure Understander processes the diagram, it creates knowledge base objects for all the diagram objects [Rajagopalan and Kuipers 94], and stores both the spatial information about the objects and the semantic information given through the picture semantics.

After processing of the diagram and the text sentence 'My name is Raman', the knowledge base will contain descriptions for all the objects in the diagram and their internal names will correspond to the abbreviations given in the picture semantics shown in Figure 4. For the reader's convenience, we show the internal names for the larger objects in the scene in Figure 2.

Let us now query the entire set of two dimensional objects in the knowledge base to identify the objects with bindings for the name-of relation.

```
algy> (ask '((member convex-2-d-objects ?x)
             (name-of ?x ?y)) :collect '(?x ?y))
  ((O RAMAN))
```

Note that the only object with a domain-specific name is the observer, O, whose name is Raman. Next, we shall proceed to process sentences 2-5 from Figure 3. Since the query forms are all similar to the one shown in Figure 5, we will not show them all here. Instead, we show only the query form for the sentence 'The name of the building that is behind Taylor Hall is the Aerospace-Engineering building.' Since the intrinsic front of Taylor Hall, given by the vector attached to object B3 in Figure 2, points to the right, the relation 'behind' is determined with respect to this new frame of reference to identify the Aerospace Engineering building, B.

```
algy> (interpret
       '(The name of the building that
         is behind Taylor Hall is the
         Aerospace-Engineering building))

 The following Algernon path shall be asserted:

((:THE ?V1 (MEMBER CONVEX-2-D-OBJECTS ?V1)
           (NAME-OF ?V1 '(TAYLOR HALL)))
 (:THE ?V2 (MEMBER BUILDINGS ?V2)
           (:BIND ?VECT
                  (GET-INT-FRONT-DIR-IF-ANY '?V1))
           (DIR-REL-WITH-FOR STRICTLY-BEHIND ?V2
                             ?V1 SCENE-INIT ?VECT))
 (NAME-OF ?V2 '(AEROSPACE-ENGINEERING BUILDING)))
```

Set of named objects after processing sentences 2-5 in Figure 3:

```
algy> (ask '((member convex-2-d-objects ?x)
             (name-of ?x ?y)) :collect '(?x ?y))
  ((B (AEROSPACE-ENGINEERING BUILDING))
   (B1 (BIOLOGY BUILDING)) (B2 (WELCH HALL))
   (B3 (TAYLOR HALL)) (O RAMAN))
```

Now let us proceed to name the streets. Below, we show the Algernon query forms generated for sentences 6 and 7 in Figure 3. Notice that there are no attempts to identify an intrinsic frame of reference since the spatial relations to be proven, enclosed-in and intersecting, are not directional relations.

```
algy> (interpret
       '(I am traveling on Twenty-Fourth street))

 The following Algernon path shall be asserted:

((:THE ?V1 (MEMBER OBSERVERS ?V1))
 (:THE ?V2 (MEMBER ROADS ?V2)
           (ENCLOSED-IN ?V1 ?V2 SCENE-INIT))
 (NAME-OF ?V2 '(TWENTY-FOURTH STREET))
 (ACTION-OF ?V1 'TRAVELING))

algy> (interpret '(Speedway street intersects
                   Twenty-Fourth street))

 The following Algernon path shall be asserted:

((:THE ?V1 (MEMBER ROADS ?V1)
           (NAME-OF ?V1 '(TWENTY-FOURTH STREET)))
 (:THE ?V2 (MEMBER ROADS ?V2)
           (INTERSECTING ?V2 ?V1 SCENE-INIT))
 (NAME-OF ?V2 '(SPEEDWAY STREET)))

algy> (ask '((member convex-2-d-objects ?x)
             (isa ?x roads)
             (name-of ?x ?y)) :collect '(?x ?y))
((R (SPEEDWAY STREET))
 (R1 (TWENTY-FOURTH STREET)))
```

6 Summary and Conclusions

As shown by the recent symposium on diagrammatic representations [SpringSymp92], there has been much interest in utilizing text and diagram input in AI applications. Novak and Bulko [90,93] have demonstrated a methodology using a blackboard architecture for utilizing text and diagrams to solve textbook physics problems. In their work, diagram input is treated as simulated visual input, consisting of interconnected lines, arcs, and points. They demonstrated effective results on several non-trivial examples, but their approach requires addressing ambiguities in the interpretation of the diagram. They suffer from the classic visual scene analysis problem [Ballard and Brown 82] of deciding which line segments belong together as parts of the same object, and which are only accidentally intersecting. Even after the diagram objects of interest are isolated, the task of matching object references in text with diagram objects is difficult since the diagram may provide few clues about the purposes of objects. PICTION [Srihari 91] is another example of a system for coordinating text and pictoral input given in visual form.

Our work complements the previously developed methods by considering the problem of integrating text and *graphically-drawn* diagrams. This medium has the advantage that a geometric representation of the objects is provided. Then, when the diagram objects are in one-to-one correspondence with the conceptual objects to be created in the knowledge base, the output of the graphical editor directly identifies the objects of interest. Further, since graphical editors provide additional methods, such as colors, filling patterns, and different brush types for highlighting objects, we can use these features to designate a semantic purpose for objects within the diagram itself.

We showed that one could then straightforwardly solve the coreference problem, particularly when each diagram object is of a unique conceptual class. For cases where the diagram contains multiple instances of objects from the same class, we showed that spatial relations could be used in the text to isolate the appropriate diagram objects. We presented our spatial representation and discussed its usefulness for modeling the orientation and relative positions of objects in both static and dynamically changing worlds.

In future work, we will address methods for reasoning about composite objects - larger objects composed of semantically meaningful subcomponents. In Figure 2, The objects marked L and L1 are both parts of a single parking lot. We have successfully implemented definitions of the form "Any two objects of the same type that are connected at an edge are parts of a single conceptual object", but are still investigating methods for reasoning about spatial relationships involving composite objects.

Some relations, such as 'equal', require an 'AND' operation that covers all the subparts. Some relations, such as 'inside' appear to require an 'OR' operation, but there may be special cases. For example, we could answer the query "Is the vehicle V inside the parking lot?" by proving that V is inside L *OR* inside L1. However, if V overlaps L and L1, this strategy is insufficient. The question "Is building B to the right of the parking lot?" is even tricker to answer, since building B is to the right of L, but is not to the right of L1.

7 Acknowledgements

I thank Professor Benjamin Kuipers for his many suggestions in developing the ideas presented in this paper.

8 References

[Abella and Kender 93] Abella, A. and J. Kender. Qualitatively Describing Objects Using Spatial Prepositions. In: Proceedings Eleventh National Conference on Artificial Intelligence (AAAI-93), 1993.

[Ballard and Brown 82] Ballard, D.H. and C.M. Brown. *Computer Vision.* Prentice Hall, 1982.

[Crawford and Kuipers 91] Crawford, J. and B. Kuipers. Algernon - A Tractable System for Knowledge-Representation. In working notes: AAAI Spring Symposium on Implemented Knowledge Representation and Reasoning Systems, 1991.

[Cui, et. al. 92] Cui, Z., A. G. Cohn, and D. Randell. Qualitative Simulation Based on a Logical Formalism of Space and Time. In: Proceedings Tenth National Conference on Artificial Intelligence (AAAI-92), 1992.

[Feiner and McKeown 90] Feiner, S. and K. McKeown. Coordinating Text and Graphics in Explanation Generation. In: Proceedings Eighth National Conference on Artificial Intelligence (AAAI-90), 1990.

[Herskovits 85] Herskovits, A. Semantics and Pragmatics of Locative Expressions. *Cognitive Science* (1985) **9**:3, 341-378.

[Larkin and Simon 87] Larkin, J. and H. Simon. Why a Diagram is (Sometimes) Worth 10,000 Words. *Cognitive Science* (1987) **11**, 65-99.

[Maass 94] Maaß, W. From Visual Perception to Multimodal Communication: Incremental Route Descriptions. *AI Review Journal* (1994) **1-3**, forthcoming.

[Mukerjee and Joe 90] Mukerjee, A. and G. Joe. A Qualitative Model for Space. In: Proceedings Eighth National Conference on Artificial Intelligence (AAAI-90), 1990.

[Novak and Bulko 90] Novak, G. S., and W. Bulko. Understanding Natural Language with Diagrams. In: Proceedings Eighth National Conference on Artificial Intelligence (AAAI-90), 1990.

[Novak and Bulko 93] Novak, G. S., and W. Bulko. Diagrams and Text as Computer Input. *Journal of Visual Languages and Computing* (1993) **4**, 161-175.

[Rajagopalan 93] Rajagopalan, R. A Model of Spatial Position Based on Extremal Points. In: Proceedings ACM Workshop on Advances in Geographic Information Systems, 1993.

[Rajagopalan 94] Rajagopalan, R. A Model for Integrated Spatial and Dynamic Reasoning about Physical Systems. In: Proceedings Twelfth National Conference on Artificial Intelligence (AAAI-94), 1994.

[Rajagopalan and Kuipers 94] Rajagopalan, R. and B. Kuipers. The Figure Understander: A System for Integrating Text and Diagram Input to a Knowledge Base. In: Proceedings Seventh International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE-94), 1994.

[Resnick and Halliday 88] Resnick J. and D. Halliday. *Fundamentals of Physics*, 3rd. Edition, John Wiley and Sons, 1988.

[Retz-Schmidt 88] Retz-Schmidt, G. Various Views on Spatial Prepositions. *AI magazine*, (1988) **9**(2).

[Spring Symp 92] Working Notes: AAAI Spring Symposium Series. Symposium: Reasoning with Diagrammatic Representations, Stanford University, 1992.

[Srihari 91] Srihari, R. PICTION: A System that Uses Captions to Label Human Faces in Newspaper Photographs. In: Proceedings Ninth National Conference on Artificial Intelligence (AAAI-91), 1991.