How Near Is Too Far? Talking about Visual Images

Uri Zernik and Barbara J. Vivier
Artificial Intelligence Program
General Electric Corporate Research and Development
Schenectady, NY 12301 USA

I. Visual Semantics Given as Search Directives

Lexical semantics must account eventually for the description of visual images. Ironically, existing linguistic systems are better geared to interfacing with abstract databases than to handling simple words such as **far** and **near**, **in** and **out**. Why are visual semantics elusive? Due to clarity of vision, flaws in semantic theories cannot pass unnoticed as they do in highly cerebral domains such as contract law or company takeovers. We have developed a theory, *directive semantics*, for dealing with visual descriptions.

This theory comes as an antithesis to pervasive theories of lexical semantics [4], which view language encoding/decoding as a process of constraint satisfaction. In the absence of a task domain, evaluation of theoretical results is by the linguist's ear: unusual utterances are "starred" out at convenience. In contrast, we view language processing as an objective oriented task, in which lexical semantics facilitate task performance. The testbed for the theory is a well-defined cognitive task, called "hide and seek." Linguistic utterances are used to lead a *seeing* agent toward an object *implicit*[1] in a scene.

Consider the following scenario, in which a tourist is introduced to an ancient spot in Florence.

(1) **This tall arch you see near the cathedral was designed by Pilucco two thousand years ago.**

[1] Implicit objects require computational evaluation for them to be recognized.

Sentence (1) is interpreted by the listener/viewer relative to the scene currently being perceived, the skyline of the city, in which the cathedral is prominent. Following the linguistic instruction, the viewer's eyes are guided to the cathedral, then to the arch, on which his attention is focused.

Computationally, linguistic clues such as **this**, **tall**, **arch**, and **near** are problematic for traditional linguistic systems. Consider **tall**, for example: is it a descriptive constraint? Is **tall** intended to disambiguate one particular arch by ruling out many other explicitly given candidates? Not according to the scenario above, in which no arch is explicitly identified by the viewer. **Tall** is provided as a *directive* for finding an as yet unidentified arch. **Tall** guides the viewer to search for an arch over the skyline of the city. Similarly, **near** is *not* intended as a quantifier on distance. In fact, as a quantifier, **near** is not operational in task performance—is **near** defined as 5 or perhaps as 55? Rather, **near** is provided as a search directive: start scanning the skyline from the cathedral and away! Thus, we discover the impact of lexical semantics not in abstraction but within a concrete visual task.

In this paper we describe a computational model which receives both textual and visual inputs; we explain why the direct linkage of words with geometrical description fails, and why there is a need for broader interpretation of visual relations; we show how an image understanding "agent" is guided by the linguistic text in search of perceived objects in a scene.

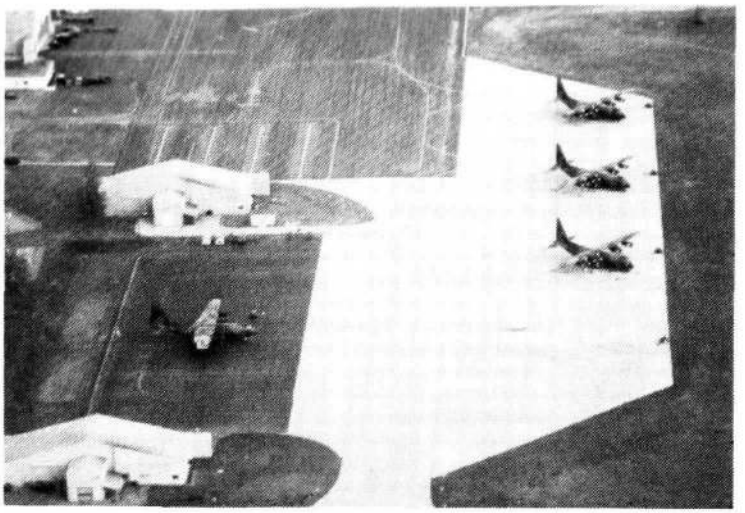


Figure 1: The Input Scene

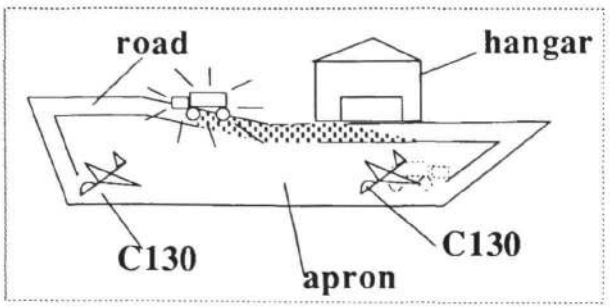
Figure 2: The Segmented Picture

A representative scene and the segmentation of the scene by the image understanding system are shown in the above figures. Figure 1 shows a photo of the Schenectady National Guard Airbase, its facilities and 3 C130 aircraft used for rescue and transportation. This picture, represented as a gray-level matrix, is converted by segmentation into a set of edges and vertices, a portion of which is shown in Figure 2. Some fixed objects in this scene are given explicitly: the hangars, the apron and the roads used by vehicles on the apron. Objects hidden (obscured by segmentation) in the picture are discovered by an image understander guided by SEER.

II. The Program SEER

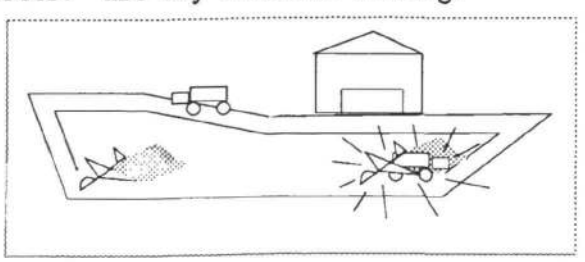
We are designing the program SEER (pronounced see'er) whose external behavior is described below.

User: **Find the fuel truck between the hangar and the C130 on the apron!**



SEER must identify a specific instance of a truck, based on references to fixed objects (i.e., the hangar, the apron). It first must find[2] the C130 aircraft, and then it searches for the truck itself based on the geometrical relation: **between the hangar and the C130**. However, **between** is not taken literally—the truck must not necessarily be on the line between the hangar and the aircraft—but it is taken as a conceptual relation. Accordingly, the search for a truck is performed along the roads connecting the two locations. As scanning of the scene progresses in search of the truck, the user receives a graphical indication of the search direction along the roads. Finally, the identified truck is illuminated.

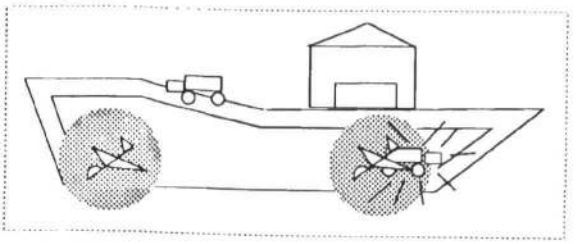
User: **Are any aircraft fueling?**



A search for a *visual script*, the C130-Fueling Script, is conducted. Accordingly, a fueling truck and a C130 in a particular spatial configuration are sought. This configuration is identified in the picture, and its constituents are marked.

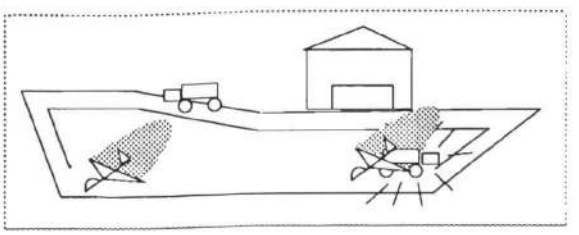
[2]In each picture, the search path is shaded; the annotations in the first drawing are for reference only.

User: **Find the truck near the aircraft.**



References such as **truck** and **aircraft** are too generic for application by an image understander. Here these references are resolved by prior *discourse*: **the aircraft** is the C130 from the first sentence, and **the truck** is resolved as either the fuel truck implicit in the fueling script or the fueling truck explicitly mentioned in the first sentence. This ambiguity is resolved by the relation **near**. The search for a truck is conducted within extending circles centered at the C130.

User: **Is the truck behind the aircraft too far?**



The relation **behind** is subjective; it must be interpreted relative to the user's *perspective*. Only by accounting for this perspective, can the program relate to the user's intended meaning. Once the truck is identified, script-based constraints are used to evaluate whether it is within the limitations of the fueling script. The last two sentences contain an apparent logical discrepancy; an instance of a truck is both "near and "too far", in the same context. This discrepancy is not real since each case has its own particular search intentions.

As demonstrated by these queries, linguistic expressions are used to direct visual scene analysis.

III. Cognitive and Technological Issues

Ultimately, cognitive models must receive and exploit input from the external world via two modalities: language and visual perception. So far, these modalities are investigated by two separate disciplines. For them to interact, a tremendous gap must be bridged between linguistic/conceptual representation of words and visual/geometrical representation of objects. In addressing this problem we choose not to ignore the main two challenges: language—*vagueness* of the medium, and vision—*brittleness* of the technology.

Consider the various interpretations of **between** in the following sequence of *locatives* [3].

(1) Point A is between points B and C.
(2) Chicago is between LA and Albany.
(3) John was watching Mary.
Barbara moved between them.
(4) My finger got stuck between the table and the wall.
(5) I find myself between a rock and a hard place.

Sentence (1), uttered by a mathematician, might possibly carry some geometrical precision. Sentence (2) uttered by an airline traveler is correct because the traveler transfers planes at Chicago en route from LA to Albany. However, it is *geometrically* incorrect: Chicago is not on the straight line constructed between LA and Albany. This discrepancy does not reflect lack of precision, which could be rectified by allowing some tolerance. It reflects the inherent complexity of linguistic relations. Sentence (3) is geometrically precise—Barbara is indeed on the line between John and Mary–but this is coincidental, due to the linear nature of light rays. The main implication of the expression is causal: Barbara "disenabled" John's seeing Mary. Sentence (4) also conveys a causal relation. Finally, sentence (5) involves two layers of metaphor: first the causal extension of **between** (X is vulnerable because he will be crushed when Y meets Z), and second, the analogy from the physical domain to another abstract domain. Thus, locatives by their nature defy direct geometrical interpretation. While allowing human speakers a great expressive power, they place on the listener the burden of identifying the appropriate implication.

Existing Image Understanding (IU) technology is quite limited. For one thing an IU program cannot identify a geometrically ill-defined object such as a person or an animal. For another, current programs cannot search for generic objects such as a cathedral or a truck. The search must be driven by a concise truck model; the slightest deviation between the designated model and the actual instance yields an identification failure. These limitations lead to three unfortunate results: (1) Without a hierarchical database (of which all models are specific instances) and gradual discrimination, all the models in the database must be exhaustively applied. All possible trucks and cathedrals are sought in each scene. (2) Without spatial relations, scenes must be scanned exhaustively. Trucks are sought in the ocean and cathedrals are sought on freeways. (3) Without conceptual expectations, there is no notion that cars drive only right-side up or that driving is oriented along the road axis. Commonsense constraints are not exploited to limit the search.

In our model, we demonstrate how these limitations of vision are overcome by linguistic inputs: (1) visual object identification is supported by linguistic reference resolution, (2) spatial relations are resolved by plan recognition, and (3) vision is driven by linguistic directives. These points are elaborated in the next sections.

IV. Resolving Locative Expressions

The resolution of locative expressions can be viewed either as a process of selection or as a process of search. We explain why, when talking about images, only the second approach is valid.

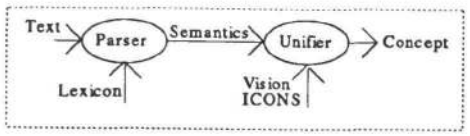
Locative Resolution as Selection: Consider, for example, the following set of expressions

[3] We define "locative" as an expression pertaining to a perceived scene. We do not use Herskovits' definition, namely, "any spatial expression involving a preposition...", since there is no method for determining whether an expression is spatial or causal.

intended to identify unambiguously objects in a database:

(1) `John Marberg.`
(2) `The ship that left Long Beach to Honolulu on March 17, 1987.`
(3) `The old chair in the corner of the dining room near the picture of my parents.`

In the traditional artificial intelligence fashion such references are resolved in two steps, as shown schematically in the figure below:



(a) Convert linguistic description to semantic description; i.e., the symbol `John Marberg` is converted to a semantic template:

(person
:first-name john
:last-name marberg
:gender male)

(b) Retrieve an instance unifying templates across a database of iconic concepts; i.e., the retrieved concept in example (1) is marberg.73 whose full description in the database is:

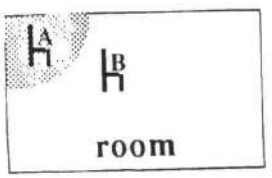
(person
:first-name john
:last-name marberg
:gender male
:age 37
:SSN 557-59-3366)

Two assumptions are made in this approach. (a) **Accessibility**: a set of concepts exists in the database in a form which lends itself to pattern matching based on features; in example (1), there is an assumed list of persons given as frame-based concepts, out of which marberg.73 is selected. (b) **Operationality**: the semantic features, given as logical predicates, are applicable in pattern matching; the strings "john" and "marberg", for example, can be validated by simple string-matching operations.

Similarly, the linguistic query of example (2) is converted to a template describing a vessel; that template is compared with all vessels in the database according to date and route; the result is a (possibly empty) set of vessels abiding by the description.

However, the two assumptions made above are not valid in certain cognitive tasks. First, as shown by Schank, Lehnert, and Kolodner[9, 1, 5], not all concepts in a semantic net are readily accessible. The main issue is not the selection of the concept but navigating in the net in order to access appropriate concepts in the first place. Second, as shown by Rosch [8], not all descriptions are crisp, and some semantic templates can be specified merely as prototypes. A prototypical dog has a tail; yet a tailless dog must still be recognized as a dog. It is yet unclear how to make prototypical descriptions operational.

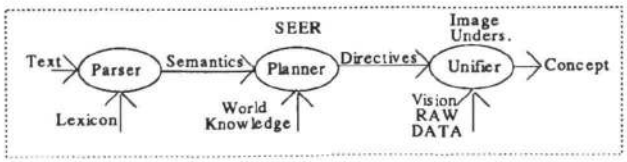
These issues are strongly manifested in the visual domain. **Accessibility**: upon receiving expression (3) above, a listener might not have accessible a list of all the chairs in the house. If really interested, the listener could walk around the house to the dining room, look at the corner and identify the chair. **Operationality**: even a simple feature such as `old` cannot become operational by a vision system. Moreover, the relation `near` is completely undefined as a quantifier. The obvious locative `in the corner` is not well-defined either as shown in the illustration below. If a corner is indeed



defined as a quarter of a cylinder, as suggested by Herskovits [4], then what is the radius of that cylinder? Is chair A "in the corner"? How about chair B?

In order to overcome these obstacles, a common practice in artificial intelligence has been to discuss "iconic vision": (a) scenes are given as collections of icons, and (b) icons are represented graphically to display the features required in pattern matching. However reality contradicts this relaxation.

Reference Identification as Search: No simplifying assumptions are made in our model about vision systems. On the contrary, we emphasize how the limitations of vision are overcome by conceptual processes. The identification task is carried out as shown below:



The scene does not contain an explicit set of icons. Mostly the scene contains raw vertices and edges. Certain objects have already been identified and precompiled. For example, the picture itself and the corners of the room have been identified in prior processing. The search is a two-step process: (a) A search planner[4] receives a sequence of search directives based on the linguistic instructions; i.e., in order to find the chair, find the dining room, the corner, the picture and then the chair. (b) The planner dispatches specific instructions to an image understander (implemented by the image understanding program [11]); i.e., `in the corner` and `near the picture` are converted into the respective directives:

```
(directive
   :prune   OUT(corner(X))
   :order   RADIAL(pos,(corner(X))))

(directive
   :order   RADIAL(pos,(picture(Y))))
```

It must still be shown how locative expressions are converted into search directives.

[4] such as that by Hanson and Riseman[3]

V. Interpreting Locative Expressions

Three approaches to the interpretation of locatives are discussed, using the examples below:

```
(1) The truck is driving between the
        hangar and the garage.
(2) The C130 in the nosedock.
```

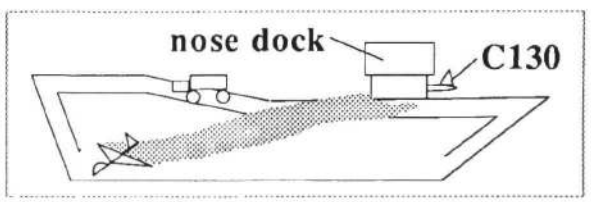
Direct Interpretation: Assuming these locatives express pure geometrical relations, a Montague-style grammar [7] can be used to convert them into geometrical primitives as in the following simplified linguistic expressions for the above examples. Thus

```
(1) A is between B and C.
(2) A is in B.
```

are converted into mathematical expressions which are based on simple geometrical primitives:

```
(1) ON(A,(CONNECT(B,C)))
(2) CONTAIN(B,A)
```

Search, following these instructions, is conducted:

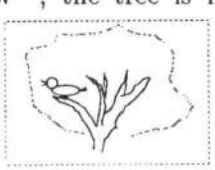


Clearly, these formulae do not capture the intended meaning: no road goes directly between B and C; the nosedock does not fully contain the C130. It is not imprecision that causes this discrepancy, but the intended meaning itself.

Geometrical Metaphor: Lakoff [6], Talmy [10], and Herskovits [4], employed geometrical metaphors in the interpretation of spatial relations. Consider, for example, the following sentence:

```
(9) The bird is in the tree.
```

As shown below [5], the tree is not really a con-



[5] the figure is taken from Herskovits' paper [4].

tainer for the bird. However, using *indirect reference* [2], the expression is reduced from 3 dimensions to 2 dimensions, meaning: the contour of the bird is in the contour of the tree, as: `CONTAIN(contour-of(tree),contour-of(bird))`

In another example, the conversion is along the whole-part dimension:

(10) John was sitting in the truck.

This example is interpreted as follows:

`CONTAIN(cabin-of(truck),john)`

This approach captures metaphors along geometrical dimensions, but it fails on metaphors which require broader world knowledge.

Object-Oriented Metaphor: The following sequence of examples shows the limitation of the geometric approach and motivates the object-oriented approach:

(10) The C130 is in the nosedock.

(11) The mouse is in the trap.

(12) John is registered in this school.

Sentence (11) can be explained by the whole-part conversion:

`CONTAIN(nosedock,part-of(C130))`

However, for a mouse's tail caught in a spring-type trap, sentence (11) is yet unexplained. The mouse is trapped, but it is not contained in the trap. Solving such relations requires knowledge about the object under analysis. The interpretation of **A is in B** depends on the nature of the interaction between A and B. Accordingly, aircraft A is **in** the nosedock B if A's nose is being maintained at dock B; mouse A is **in** trap B if A is trapped in B; person A is **in** school B if A is on the list of students of B.

The interpretation of **between** is further complicated by the *time-space duality* [6]. **The truck is between B and C** implies that at a certain point in time between leaving A and arriving at B, the truck passes through it current location. This interpretation relies on the identification of a truck as a moving vehicle, whose possible location depends on its type: aircrafts are sought on air corridors, cars are anticipated on roads, and a train is probably on a railroad. Elementary world knowledge of this kind is essential to support a visual system in finding objects.

VI. Conclusions

We have examined the interpretation of locatives within a concrete task domain. In this task we have identified two central issues. First, locatives do not possess narrow geometrical interpretation, but they require world knowledge. Second, locatives are not used as simple pointers to objects—and as we have shown, their semantics is unclear when thought of as such—but they are used as directives for navigating in visual scenes. Our theory of *directive semantics* is employed in the integration of natural language and vision.

References

[1] M.G. Dyer. *In-Depth Understanding.* MIT Press, 1983.

[2] G. Fauconnier. *Mental Spaces.* MIT Press, 1988.

[3] A.R. Hanson and E.M. Riseman. Visions: a computer system for interpreting scenes. In *Computer Vision Systems*, Academic Press, 1978.

[4] A. Herskovits. Semantics and pragmatics of locative expressions. *Cognitive Science*, 1985.

[5] J. Kolodner. *Retrieval and Organizational Strategies in Conceptual Memory.* Lawrence Erlbaum, Hillsdale, NJ, 1984.

[6] G. Lakoff and D. Johnson. *Metaphors we Live By.* Univ. of Chicago, 1980.

[7] R. Montague. On the proper treatment of quantification in ordinary english. In J. Hintikka et al., editor, *Approaches to Natural Language*, 1973.

[8] E. Rosch. Principles of categorization. In B. Lloyd, editor, *Cognition and Categorization*, Lawrence Erlbaum, 1978.

[9] R.C. Schank and R. Abelson. *Scripts, Plans, Goals, and Understanding.* Lawrence Erlbaum, Halsted, NJ, 1977.

[10] L. Talmy. How language structures space. In H. Pick et al., editor, *Spatial Orientation*, Plenum Press, 1983.

[11] D. Thompson and J.L. Mundy. Three dimensional model matching from an unconstrained viewpoint. In *Proc. IEEE Robotics and Automation Conf.*, 1987.