Visual Semantics: Extracting Visual Information from Text Accompanying Pictures *

Rohini K. Srihari and Debra T. Burhans
CEDAR/SUNY at Buffalo
UB Commons, 520 Lee Entrance- Suite 202
Buffalo, NY 14228-2567 USA
rohini@cs.buffalo.edu
burhans@cs.buffalo.edu

Abstract

This research explores the interaction of textual and photographic information in document understanding. The problem of performing general-purpose vision without a priori knowledge is difficult at best. The use of collateral information in scene understanding has been explored in computer vision systems that use scene context in the task of object identification. The work described here extends this notion by defining *visual semantics*, a theory of systematically extracting picture-specific information from text accompanying a photograph. Specifically, this paper discusses the multi-stage processing of textual captions with the following objectives: (i) predicting which objects (implicitly or explicitly mentioned in the caption) are present in the picture and (ii) generating constraints useful in locating/identifying these objects. The implementation and use of a lexicon specifically designed for the integration of linguistic and visual information is discussed. Finally, the research described here has been successfully incorporated into *PICTION*, a caption-based face identification system.

Introduction

The general problem being investigated is that of establishing a correspondence between words and the visual depictions they evoke. This correspondence is in general not one-to-one, but for certain specialized domains it is possible to establish a direct correspondence between words and the pictorial elements being referenced by them (e.g. weather maps: the word 'pressure' in a caption implies the presence of isobars in the accompanying picture). (Jackendoff 1987) attempts to establish a correspondence between words and 3D models of objects, but the problem is handled primarily at the single-word level (nouns and verbs): this work does not extend to establishing a correspondence between a sentence/phrase and the complex scene it may evoke.

*This work was supported in part by a grant from ARPA (ARPA 93-F148900-000)

In the present research, we focus on captioned pictures. Specifically, we address the problems of (i) identifying useful information in the text and (ii) extracting and representing this information so it can be used to direct a computer vision system in the task of picture understanding.

In a computer vision system, "visual information" refers to knowledge about objects that is required to detect them in a scene. This includes descriptions of objects in terms of their components and the spatial constraints between them, as well as typical scene information that relates objects in a common context. Visual information is generally represented statically, limiting the range of contexts in which it is applicable. An example is the modeling of a typical neighborhood scene comprised of streets, houses, trees, etc. (Weymouth 1986). Visual semantics plays a key role in allowing scene descriptions to be *dynamically constructed* from descriptive text. These scene descriptions can then be used by a vision system to guide knowledge-based interpretation of the associated picture.

Captions associated with two-dimensional images represent a domain that imposes a reference point on the visual image evoked by a phrase. For example, the phrase "President Clinton greets Vice President Gore" suggests Clinton and Gore are facing each other: their faces are expected to be in profile in the picture. We present a new *visual semantics* for this domain. It includes the following elements, all of which are described in this paper:

- A lexicon for integrating linguistic and visual information.
- The representation of visual information as a set of constraints that can be applied to the picture. (Strat & Fischler 1991) discusses the use of context in visual processing but does not cover the generation of constraints.
- A systematic procedure for processing a caption to generate visual constraints.

A system, *PICTION*, based on visual semantics is described. It uses information obtained from a news-

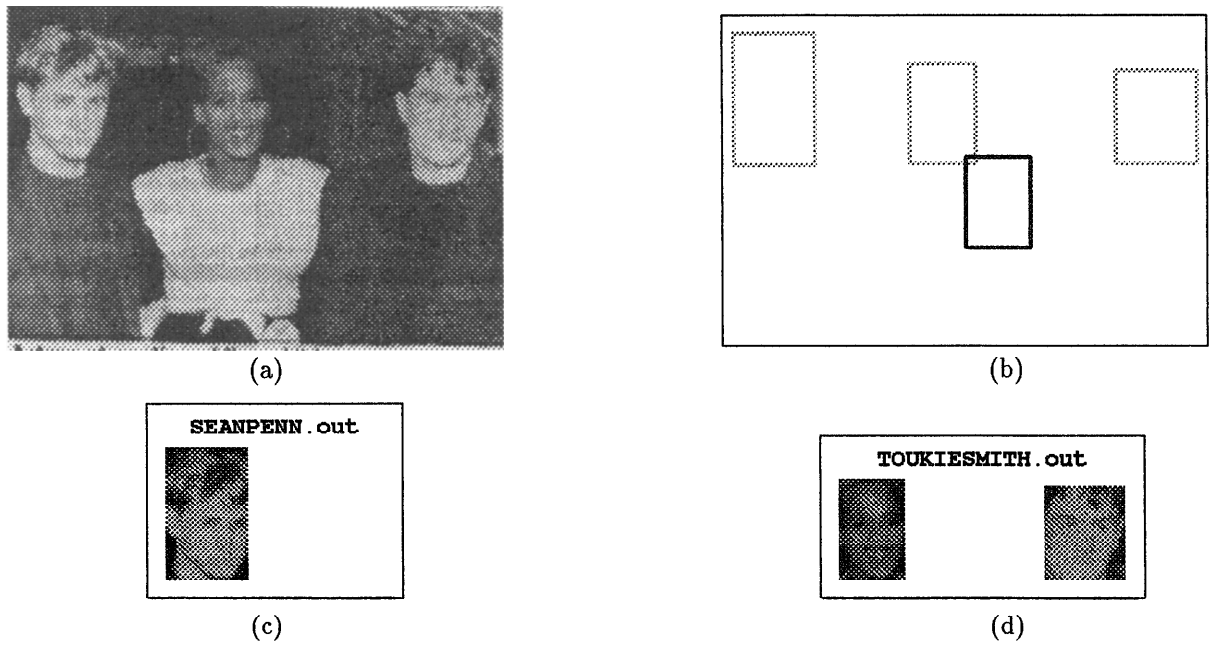

Figure 1: (a) photograph with caption "Actors Sean Penn, left, and Robert DeNiro pose with Toukie Smith, sister of the late fashion designer Willi Smith, at a New York celebrity auction Sunday in memory of Smith" (*The Buffalo News*, Feb. 27, 1989); (b) output of face locator; (c,d) output of *PICTION*.

paper caption to label faces in the accompanying photograph.

This research is most relevant in the context of document image understanding. Pictures with captions are ubiquitous in documents, newspapers and magazines. The information contained in both pictures and captions enhances overall understanding of the accompanying text, and often contributes additional information not specifically contained in the text. This information could subsequently be incorporated into an integrated text and picture database that permits *content-based retrieval.*

PICTION: A Caption-based Face Identification System

We refer to a caption and its associated picture in a newspaper as a *communicative unit.* Given a text file corresponding to a newspaper caption and a digitized version of the associated photograph, *PICTION* (Srihari 1994; 1991) is able to locate, label, and give information about objects referred to in the communicative unit. *PICTION* was initially tested on a database of 50 pictures. It successfully and uniquely identified faces in 62% of the cases, and achieved partial success on an addition 11% of the pictures.

PICTION provides a computationally less expensive alternative to traditional methods of face recognition. These methods employ model-matching techniques: only people for whom pre-stored face models exist can be identified. In *PICTION* faces are identified based solely on visual information conveyed by accompanying text. A key component of *PICTION* is the face locator, which locates (but cannot recognize) human faces in photographs.

Figure 1 is an example of a digitized newspaper photograph and accompanying caption that the system successfully processes. The male/female filter is not able to distinguish between Toukie Smith and Robert DeNiro, leading to multiple possible bindings. Sean Penn is identified correctly based on spatial constraints.

Figure 2 shows the overall control structure of *PICTION*. The three main components of *PICTION*'s architecture are (i) a natural-language processing (NLP) module, (ii) an image understanding (IU) module, and (iii) a language-image interface (LII). The NLP and IU modules interact through the LII which maintains the long-term knowledge base (LTM). *PICTION* runs on a Sun Sparcstation. It has been implemented primarily in LOOM(ISX 1991), an environment for constructing knowledge based systems, with a LISP interface to visual routines written in C.

The NLP module (illustrated in Figure 2) is divided into three stages: syntactic parsing, partial semantic interpretation (PSI) and caption based constraint generation (CBCG). The input to the NLP Module is the original newspaper caption; the output is a set of visual constraints. The LII module converts the visual information into a series of directives for the IU module

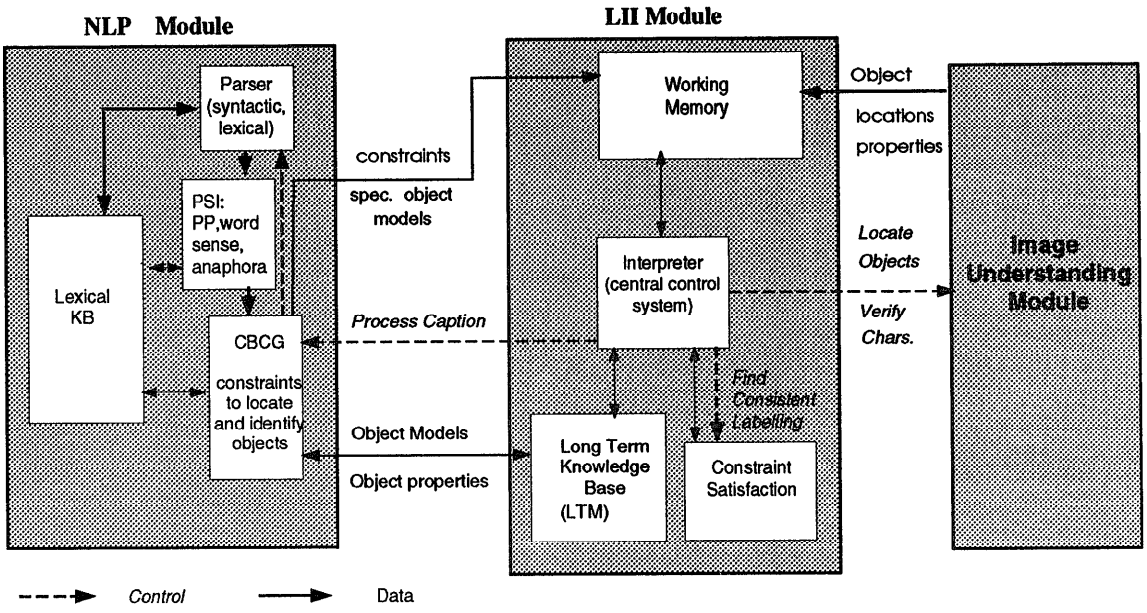

Figure 2: *PICTION* System Overview

which is then called on to interpret the picture. The IU module has several features which enable it to be guided by the LII, including: constrained search, ability to characterize objects, crude and refined object-location procedures, ability to change parameters and repeat actions, output compatibility with intermediate representation, and the ability to perform bottom-up interpretation when necessary. Information is consolidated in the LII by satisfying various types of constraints. This may require repeated calls to the IU module.

The remainder of this paper focuses on the NLP module, namely, the processing of the caption.

Lexical Database

PICTION uses a broad coverage syntactic/semantic lexical database which has been constructed from the following sources: (i) Longman Dictionary of Contemporary English (LDOCE), (ii) Oxford Advanced Learner's Dictionary (OALD), (iii) WordNet (Beckwith *et al.* 1991) (iv) name lists [1] and (v) manual augmentation (ontological information and visual semantics). This is similar to work done on the Penman Project at ISI (Knight & Luk 1994).

A LOOM knowledge base instance is constructed for each word in the lexicon. Name instances contain information on gender; word instances contain syntactic and semantic information from LDOCE (including subject field codes, semantic restrictions and verb subcategorization), morphological information from OALD and pointers to concepts representing WordNet synsets. WordNet synsets are represented as LOOM concepts organized in a hierarchy that mirrors WordNet. Synset concepts contain part-of, is-part and verb classification information from WordNet. Visual and procedural semantic properties have been added manually to a subset of the concepts.

Our approach uses and extends the *Naive Semantics* (Dahlgren 1988) model (a theory of associating commonsense knowledge with words). An upper level ontology that is based on *Naive Semantics* and extends the WordNet hierarchy is incorporated into the lexicon.

It is necessary to represent fixed, visual properties of objects, such as size and color, as well as procedural information for certain words and phrases. For example, a recent caption identified one of the people in the corresponding photograph as "the person wearing the hat". This should generate a call to an object finder with the location "above head", and the scale and shape properties of the object (hat).

|HAT| is the name of the lexical instance for the word "hat" (Figure 3). It contains information from the LDOCE and a pointer to the corresponding WordNet synset concept. "WNN0206338" represents information from WordNet. Synonyms for hat include "chapeau" and "lid"; "WNN02066195" is the superconcept of "hat". The "scale", "shape" and "function" slots have been manually instantiated and are used to generate visual information. Procedural information stored with the "wear" synset specifies that if the event "wear" is associated with "hat", and the subject of "wear" is a human, a typical location for hat is on the head of the associated person. The CBCG uses this information to generate a locative constraint (de-

[1] These lists were obtained from the Consortium for Lexical Research at New Mexico State University.

```
(tellm (:about |HAT|
 :is-primitive LDOCEconcept
  (WNsynset WNN0206338)
  (noun-part LD0030091)))

(defconcept WNN0206338
 :is-primitive WNN02066195
 :annotations
  ((word-list (hat chapeau lid))
   (scale s2)
   ($semantic-feature $clothing)
   (has-part (brim crown hatband))
   (shape
     (procedure
       (find-shape (crown,cylinder,hollow))
       (find-shape (brim,disc))
       (top-of(crown,brim))))
   (function
     (wear(E,noun,Y) & human(Y)
      & typ-location(E,noun,
      procedure(locate-in-vicinity
       (noun,top-of(locate-part(head(Y)))))))))
```

Figure 3: Partial lexicon generation code for the word "hat"

scribed later) which is subsequently used by the LII and the vision module to identify the person wearing the hat.

Visual Hierarchies

We have defined visual hierarchies in terms of *visual superconcepts* which reflect type (man-made, natural), shape, texture properties, boundary properties etc. of an object. New links (*visual-is-a, visual-part-of*) have been added between existing WordNet synsets (representing concrete objects) and these superconcepts. Specialized attributes such as size, color, expected location, etc. are added at the synset level. This visual information allows recognition tools (such as segmentation tools, edge detectors, surface detectors) and specialized object detectors for certain common object classes (e.g., human face, tree, building, car) to be appropriately invoked. [2]

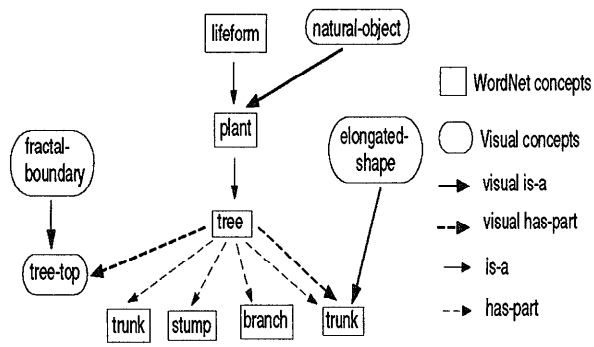

Figure 4: Visual hierarchy superimposed on WordNet concept hierarchy.

There are many objects for which it is difficult to construct detailed shape descriptions. In these cases it is sufficient to state some of their properties such as natural/man-made, boundary description, etc. Identification of these objects is based on a *blob* theory of object recognition: using constraints that specify size, expected location and a few object properties, the object can be roughly located, which is sufficient for our purposes.

For example, consider a picture of a man holding a trophy accompanied by the caption "Thomas Smith holding the trophy he won at ...". To identify the trophy, the system would first find the face of Thomas Smith, then search in the appropriate vicinity for an object exhibiting the required properties (man-made, small-medium size, etc.).

Our representation allows objects to be modeled at various resolutions. For example, at the most general level, a tree is a natural object with a fractal boundary. A more detailed visual model of tree defines the visual parts of a tree as well as the spatial relationships between these parts. These parts are classified according to the chosen set of visual superconcepts. At the most specific level, the description of a tree includes a specialized recognition module for trees.

The has-part information in visual object models may differ from the has-part information used in WordNet for the following reasons: (i) the has-part information in WordNet may be too fine-grained to be exploited by a vision system (e.g., shoes can have laces), and (ii) the names of the parts may differ (e.g., the WordNet entry for 'tree' includes crown, trunk, branches and stump, however the visual description for tree will have a treetop and a trunk). This is illustrated in Figure 4.

A study in human cognition regarding the task of identifying objects reveals that there is justification for having different abstractions for words and pictures. (Linde 1982) postulates the existence of two separate semantic-memory representations. (Jolicoeur, Gluck, & Kosslyn 1984) states that both words and pictures may use the same semantic-memory representation; they differ however in the *entry point*, namely the particular level of abstraction in the hierarchy at which the association is made.

Parsing

Pre-Processing Input

There are two objectives to this phase. The first is the elimination of *directive* phrases such as "left-of", "front row, left to right", etc. Directive phrases are associated with appropriate noun phrases and passed directly to

[2] According to (Biederman 1988), there are about 3000 common entry-level objects which the human perceptual system can detect.

the CBCG for further processing. The second objective of pre-processing is the detection and classification of proper noun sequences, which frequently contain unknown words. In this example from *The Buffalo News*, the proper noun sequences are bracketed:
`[Vladimir Horowitz] at [Steinway and Sons], [New York],...`
"Vladimir Horowitz" is classified as a name, "Steinway and Sons" as an organization name and "New York" as a geographic location.

A hidden Markov model, trained on a portion of the Penn Treebank corpus, is used to detect proper noun sequences. We are currently able to detect proper noun sequences with about 90% accuracy. Proper noun lists containing appropriate classification information (gender, location, organization, etc.) [3] and heuristic rules (involving typical suffixes such as "Inc.") are employed in the categorization of proper nouns. Correct classification of proper noun sequences still poses a significant challenge (Mani *et al.* 1993).

LFG Parser and Partial Semantic Interpretation

An LFG which covers basic features of English grammar has been compiled into an LR parsing table. We employ an efficient LR parser augmented by pseudo/full unification packages (Tomita 1987). Certain semantic features (e.g., animate/inanimate; features associated with proper nouns) are incorporated into the output structure of the parser. This semantic information is obtained solely through lexicon lookup and is used to help disambiguate among multiple syntactic structures output by the parser. We are experimenting with statistical techniques for handling prepositional phrase and word sense disambiguation. Anaphoric references are resolved by the use of weights assigned to various referents depending on their role in the sentence.

Caption-based Constraint Generator (CBCG)

Input to the CBCG is the disambiguated parse from the PSI stage; output is a set of visual constraints to be used by the IU module. The CBCG makes use of the LTM in the LII for retrieval of information about well-known people, as well as the lexicon, for assigning visual semantics to the parse.

Constraint Types

Using visual information derived from text, *PICTION* hypothesizes a set of objects expected to be in the picture and constraints on those objects. Constraints are divided into four types:

- **Spatial Constraints** are geometric or topological constraints, such as left-of, above, inside, etc. They can be binary or n-ary, and describe inter-object relationships. Complex spatial constraints such as "surround" are broken down into a set of constraints based on spatial primitives.
- **Locative constraints** express information about the location of objects in the picture with respect to a particular frame of reference. The information conveyed is procedural in nature, for example, if you are told there is a chair in the corner, it results in the following high-level procedure construct: $loc\_in\_vicin(chair, region(corner(entire\_image)))$.
- **Characteristic Constraints** are unary constraints which describe properties of objects. Examples include gender and hair color.
- **Contextual Constraints** are those which describe the setting of the picture, and the objects which are expected to appear. Examples include the people present (mentioned in the caption), whether it is an outdoor scene, and general scene context (apartment, airport, etc).

Some visual constraints are expressed explicitly as assertions, for example *left_of(person_a person_b)*. Locative and characteristic constraints are implicit in the object model. Contextual constraints consist of the instantiated objects and an asserted general scene context.

Consider the photograph and caption of Figure 1. Some of the constraints output by the CBCG for this example are:

```
SPATIAL: left_of(Sean Penn ,Robert DeNiro)
         adjacent(Robert DeNiro,Toukie Smith)
CHARACT: has_prop(name:Sean Penn;gender:male)
         has_prop(name:Toukie Smith;gender:female)
```

"Adjacent" is asserted as the default spatial constraint.

Automatic Generation of Visual Constraints

The CBCG has been written as a rule-based system which uses LOOM concepts to drive the semantic "parsing". LOOM methods and rules are invoked when particular concepts are instantiated.

There are three main categories of rules.

- **Word-based:**
  Spatial and characteristic constraints are frequently indicated by single words. Examples include left, right, above and below, as well as characteristics such as hair color and titles like President.
- **Phrase-based:**
  Locative and characteristic constraints are often indicated by directive phrases. Examples include "between the two buildings" and "wearing the striped shirt".
- **Sentence-based:**
  Contextual constraints can generally be inferred at the sentence level, taking into account the various

[3] These lists were obtained from the Consortium for Lexical Research at New Mexico State University.

objects mentioned and their relations and properties. An example of this is the "SVOPP" rule which states that if the sentence is of the form subject-verb-object-prepositional_phrase, and both the subject and object represent humans, and the PP represents a time and/or location, then propose that both the subject and the object are in the picture.

The top-level rules are based on syntactic structure and attempt to predict which people (or objects) are in the picture. Verifying the antecedents of these rules (e.g., is the person deceased?) causes other rules to be fired. The final action is to generate identifying information for every person/object predicted to be in the picture.

Consider the caption "Actors Sean Penn, left, and Robert DeNiro pose with Toukie Smith, sister of the late fashion designer Willi Smith, at a New York celebrity auction Sunday". All three people mentioned will be predicted to be in the picture as the sentence has the form <subject-list> <verb> <object> <adverbial place> <adverbial time>.

A concept and the associated production rule used to generate the contextual constraint that the picture is indoors are as follows:

```
(defproduction is-indoor
 :when (:detects (inside ?parse ?cg))
 :perform (tell (:about ?cg (location 'indoor))))
(defconcept inside
 :is (:predicate (?parse)
   (let ((?flat (explode ?parse))
         (semfeats nil))
         (dolist (?x ?flat)
           (push (get-sem-features ?x) ?semfeats))
         (or (memberp '$indoor ?semfeats)
           (and (memberp '$social-event $semfeats)
            (not (memberp $outdoor ?semfeats)))))))
```

The photograph in Figure 1 is predicted to be of an indoor scene since "auction" has the semantic feature `$social-event` stored in the lexicon, and there is nothing mentioned which has the semantic feature of being outdoors. The fact that Toukie Smith is female is inferred by the presence of the word "sister" in the directional phrase associated with her name. The spatial constraints shown in the previous section are generated when spatial constraint rules are fired.

Summary

This paper has presented a new theory of visual semantics that concerns the use of descriptive text in the interpretation of accompanying photographs. Although the examples used are from captioned newspaper photographs, and the application is knowledge-based vision, this work can be extended to any domain where both language and pictures are used to communicate information. A highlight of this work is the development of a lexicon that includes visual hierarchies, as well as a systematic procedure for generating visual constraints from text accompanying a picture.

References

Beckwith, R.; Fellbaum, C.; Gross, D.; and Miller, G. A. 1991. WordNet: A Lexical Database Organized on Psycholinguistic Principles. In *Lexicons: Using On-line Resources to Build a Lexicon.* Lawrence Erlbaum.

Biederman, I. 1988. Aspects and extensions of a theory of human image understanding. In Pylyshyn, Z., ed., *Computational Processes in Human Vision: An interdisciplinary perspective.* Ablex.

Dahlgren, K. 1988. *Naive Semantics for Natural Language Understanding.* Boston: Kluwer Academic Press.

ISX Corporation. 1991. *LOOM Users Guide, Version 1.4.*

Jackendoff, R. 1987. On Beyond Zebra: The Relation of Linguistic and Visual Information. *Cognition* 26(2):89–114.

Jolicoeur, P.; Gluck, M. A.; and Kosslyn, S. M. 1984. Pictures and Names: Making the Connection. *Cognitive Psychology* 16:243–275.

Knight, K., and Luk, S. 1994. Building a Large Scale Knowledge Base for Machine Translation. Forthcoming. In *Proceedings of AAAI-94.*

Linde, D. J. 1982. Picture-word differences in decision latency. *Journal of Experimental Psychology: Learning, Memory and Cognition* 8:584–598.

Mani, I.; MacMillan, T. R.; Luperfoy, S.; Lusher, E. P.; and Laskowski, S. J. 1993. Identifying Unknown Proper Names in Newswire Text. In *Proceedings of the Workshop on Acquisition of Lexical Knowledge from Text,* 44–54.

Srihari, R. K. 1991. PICTION: A System that Uses Captions to Label Human Faces in Newspaper Photographs. In *Proceedings of AAAI-91,* 80–85. AAAI Press.

Srihari, R. K. 1994. Use of Collateral Text in Understanding Photos. Forthcoming. *Artificial Intelligence Review.* Special Issue on Integration of NLP and Vision.

Strat, T. M., and Fischler, M. A. 1991. Context-Based Vision: Recognizing Objects Using Information from Both 2-D and 3-D Imagery. *IEEE PAMI* 13(10):1050–1065.

Tomita, M. 1987. An Efficient Augmented-Context-Free Parsing Algorithm. *Computational Linguistics* 13(1-2):31–46.

Weymouth, T. 1986. *Using Object Descriptions in a Schema Network for Machine Vision.* Ph.D. Dissertation, University of Masschussetts at Amherst.