Control Structures for Incorporating Picture-Specific Context in Image Interpretation

Rajiv Chopra
Department of Computer Science
226 Bell Hall
State University of New York at Buffalo
Buffalo, New York 14260
U. S. A.

Rohini K. Srihari*
Center for Document Analysis and Recognition
520 Lee Entrance- Suite 202
State University of New York at Buffalo
Buffalo, New York 14228-2567
U. S. A.

Abstract

This paper describes an efficient control mechanism for incorporating picture-specific context in the task of image interpretation. Although other knowledge-based vision systems use general domain context in reducing the computational burden of image interpretation, to our knowledge, this is the first effort in exploring picture-specific collateral information. We assume that constraints on the picture are generated from a natural language understanding module which processes descriptive text accompanying the pictures. We have developed a unified framework for exploiting these constraints both in the object location and identification (labeling) stage. In particular, we describe a technique for incorporating constrained search in context-based vision. Finally, we demonstrate the effectiveness of this approach in PICTION, a system that uses captions to label human faces in newspaper photographs.

1 Introduction

To solve the inherently under-constrained task of image interpretation, additional sources of knowledge have successively been added to provide the necessary constraints for the recognition of objects and understanding of scenes.

Vision systems have attempted to exploit general knowledge assumed applicable to all scenes and specific knowledge of scene objects and their appearance/configuration (model based vision). As a result, complex control strategies to utilize knowledge in the image interpretation task have been suggested. Devising control structures for knowledge based vision has been an active area of research since the mid 1960's with several refinements of the *cycle of perception* and the development of hierarchical, heterarchical and blackboard models. See [Firschein and Fischler, 1987; Chellapa and Kashyap, 1992] for the historical development of control structures in image understanding.

*This work was supported in part by ARPA Contract 93-F148900-000.

There has also been recent work in domain-independent control structures where interpretation can be formally stated based on the semantics of a standardized representation[Niemann *et al.*, 1990].

1.1 Collateral Based Vision

Our research has focused on developing a computational model for "understanding" pictures based on accompanying, descriptive text. Understanding a picture can be informally defined as the goal-driven process of locating and identifying objects whose presence in the image is suggested by the accompanying text. The ideas presented in this paper carry the notion of top-down control one step further, incorporating not only general context but a high-level description of the actual picture.

We have divided this problem into two major subtasks. The first task is the processing of language input. We are developing a theory called *visual semantics* which describes a systematic method for extracting and representing useful information from text pertaining to an accompanying picture. This information is represented as a set of constraints. The processing of captions for generating constraints is discussed in [Srihari and Burhans, 1994]. The second major task involves the design of an architecture wherein collateral information from text can be efficiently exploited in the process of image interpretation. This second subtask is the focus of this paper.

There are several applications which are enabled by collateral-based vision. These include diagram understanding, indexing and retrieval from text/image databases and aerial image analysis.

We first present an overview of PICTION, a caption based face identification system. Section 3 discusses control in Piction, constraint satisfaction as a model for control, spatial reasoning and a general algorithm which exploits constrained search. The next section highlights some limitations and discusses future work to overcome them.

2 PICTION

The computational model for collateral-based vision we have developed can be applied to any situation where photos are accompanied by descriptive text. In describing the model however, it is useful to present working examples based on this model. For this reason, we refer

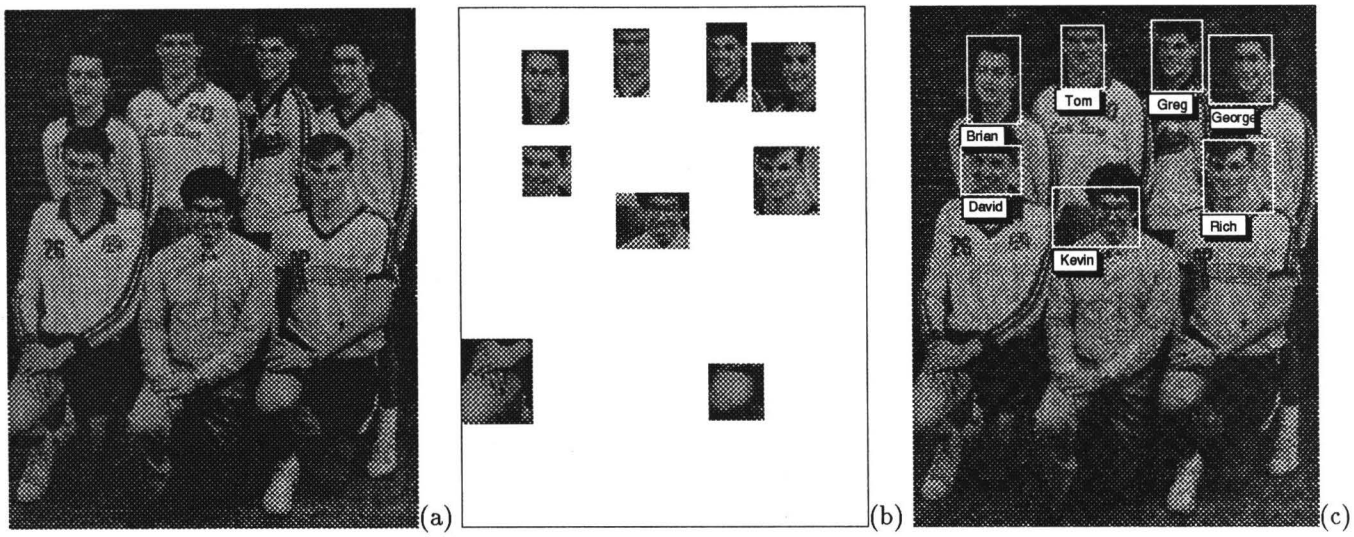

Figure 1: **Caption:**"The All-WNY boys volleyball team surrounds the coach of the year, Frontier's Kevin Starr. Top row, from left, are Eden's Brian Heffernan, Lake Shore's Tom Kait, Sweet Home's Greg Gegenfurtner, Kenmore West's George Beiter. Bottom row, from left, are Eden's David Nagowski and Orchard Park's Rich Bland."
(a) Original image (b) Output of face locator (c) Labeled faces

to the system *PICTION* [Srihari *et al.*, 1994], that identifies human faces in newspaper photographs based on information contained in the associated caption. PICTION is noteworthy since it provides a computationally less expensive alternative to traditional methods of face recognition in situations where pictures are accompanied by descriptive text. Traditional methods employ model-matching techniques and thus require that face models be present for all people, for them to be identified by the system.

```
CONTEXTUAL:
    in-picture(name: Kevin Starr; class: human)
    in-picture(name: Brian Heffernan; class: human)

CHARACTERISTIC:
    gender(name: Brian Heffernan, male)
    gender(name: Tom Kait, male)

SPATIAL:
    surrounded-by(Kevin Starr,
                  (Brian Heffernan, Tom Kait ...))
    strict-left-of(Brian Heffernan, Tom Kait)
    below(David Nagowski, Brian Heffernan)
    in-row(number: 4;
           names: (Brian Heffernan, Tom Kait ... ))
    in-row(number: 2;
           names: (David Nagowski, Rich Bland ))
```

Figure 2: A subset of the constraints generated from the caption of Figure 1.

The four main components of the model are as follows: (i) a natural-language processing (NLP) module, (ii) an image understanding (IU) module, (iii) a high level control module, and (iv) an integrated language/vision knowledge base.

The NLP module processes the caption text and expresses the result as a set of constraints. **Spatial Constraints** are geometric or topological constraints, such as left-of, above, between, inside, etc. They can be binary or n-ary, and describe inter-object relationships. They are typically used to identify/disambiguate among objects but can also be used to locate them. **Characteristic Constraints** are those which describe properties of objects. Examples include gender and hair color. Finally, **Contextual Constraints** are those which describe the setting of the picture, and the predicted objects in it. For example, the people present (explicitly mentioned in the caption), whether it is an outdoor scene, and general scene context (apartment, airport, etc).

The IU module performs two basic functions, the location and segmentation of objects, and the extraction of visual properties. The face-location module[Govindaraju *et al.*, 1992] is an object locator which uses edge contours as basic features, and a "springs and templates" model in a multi-resolution framework. We are currently improving a simple neural network based gender discriminator. A line detector [Burns *et al.*, 1984] is being used to construct simple detectors for 2D rectilinear objects.

Figure 1 is an example of a digitized newspaper photograph and the accompanying caption. The task is to correctly identify each of the seven people mentioned in the caption. Information from the caption (in the form of a constraint graph) is used to locate and label faces. Figure 2 illustrates some of the constraints which were derived from the caption. The number of faces and the fact that they are aligned in rows is used by a program which locates faces in the image. Identification of the faces revolved around the correct interpretation of the word 'surround'. We have interpreted it as an n-ary constraint involving minimum distances from a given candidate to the rest. The simultaneous satisfaction of this and the remaining constraints resulted in the final,

correct labeling.

PICTION has been implemented in LOOM[ISX, 1991]. The visual routines have been implemented in C; the system runs on a Sun Sparcstation.

3 Control in PICTION

Recent work by Strat and Fischler[1991] discusses the use of context in visual processing, but focuses on the exploitation of low-level collateral information (e.g., lighting conditions). Furthermore, they do not address the generation of these constraints. We present a control strategy which exploits (i) a confident hypothesis of the image contents and (ii) all levels of contextual information to aid the visual processing.

Our approach is to select a set of object classes of interest which we would like to locate and identify in images. For PICTION this set has been restricted to one object class, the human face. However, recognizing additional classes of objects (like hats and boxes) may greatly assist us in identifying faces.

Constraint satisfaction provides a single framework for incorporating spatial, characteristic, contextual and other general domain constraints without having to overspecify control information. However the location of objects is still performed at the IU level thereby allowing existing object detectors to be integrated into the overall model.

3.1 The Single Object Class Problem: Labeling Faces

For the single object class version of PICTION, the overall control can be summarized in a few steps :

- Constraint Graph Generation (hypothesis generation: NLP module)
- Face Candidate Generation (candidate generation : IU module)
- Consistent Labeling (verification and recognition)

Knowledge of the domain allows us to make simplifying assumptions for spatial reasoning. The assumptions concern spatial representations of objects, frames of reference, scale and rotation invariance. Spatial reasoning in this domain involves verification of predicates corresponding to spatial relations. [Burhans *et al.*, 1995]

3.2 Constraint Satisfaction as a Control Paradigm

A *static* constraint satisfaction problem (V, D, C, R) involves a set of n **variables**, $V = \{v_1, v_2, \ldots v_n\}$, each with an associated **domain** D_i of possible **values**. The search space consists of the Cartesian product of the variables' domains, $D = D_1 \times D_2 \times \ldots \times D_n$. A set of **constraints** C is specified over some subsets of these variables, where each constraint C_p involves a subset $\{i_1, \ldots, i_q\}$ of V and is labeled by a relation R_p of R, subset of the Cartesian product $D_{i_1} \times \ldots \times D_{i_q}$ that specifies which values of the variables are compatible with each other. A **solution** to a CSP consists of an assignment of values to variables such that all constraints are satisfied. Various approaches exist for solving CSP's. A common approach is based on problem reduction strategies, and involves achieving local, i.e. node and arc, consistency[Mohr and Henderson, 1986; Mohr and Masini, 1988]. This reduces the amount of backtracking required for global consistency.

In PICTION, each variable v_i corresponds to an object or person hypothesis, i.e., object or person predicted to be in the picture by the constraint generator. The domain D_i for each variable is initialized to the set of all **candidates** $\{c_{i_1}, \ldots, c_{i_m}\}$ generated by a call to the face locator.

There are 3 types of constraints used in the labeling process. **Unary Constraints** typically correspond to characteristic constraints (e.g., gender, color of hair, etc.). **Binary Constraints** typically correspond to spatial constraints (e.g. left-of, above, etc.). ***N*-ary Constraints** correspond to rules obtained from the domain *a priori* but applied in a discretionary manner by the constraint satisfier, or n-ary spatial constraints like the 3-ary constraint *between*. An example of a domain dependent n-ary constraint is one which examines height relationships between people in the same row. It favors those solutions where the vertical positions of faces do not differ significantly and where there is a minimal amount of horizontal overlap.

Constraint satisfaction in a schema-based methodology has been used in computational vision, eg. the MAPSEE system[Mulder *et al.*, 1988]. Their approach is that of data driven scene labeling where image features are nodes and possible interpretations form the sets of labels. Our application of constraint satisfaction is in a goal-driven image interpretation system where a strong scene hypothesis forms the set of nodes, and located objects (features) form the sets of labels. The constraints in our system are picture specific as well as domain dependent.

3.3 HyperArc Consistency and Expensive Constraints

There are two key differences between the statement of PICTION's constraint satisfaction problem and a traditional CSP. Traditionally unary constraints are satisfied before higher order constraints. Knowing that certain binary constraints are cheaper to evaluate than unary constraints motivated a new algorithm where constraints are applied based on cost and reliability measures. Additionally, the algorithm can handle n-ary constraints without explicitly computing the complete corresponding n-ary relation [Chopra *et al.*, 1995].

3.4 Multiple Object Classes: Generalizing PICTION

So far we have seen a control strategy that exploits collateral information to identify objects of a single class in an image. This needs to be extended to more general cases involving multiple object classes. Apart from making it applicable to domains with more than one object class of interest, this capability may be required even if there is a single object class of interest. We borrow here the terminology of defining a *target* vocabulary and a *recognition* vocabulary[Strat, 1992]. Even if the target

vocabulary consists of a single object class, its recognition is greatly enhanced by the development of a large recognition vocabulary. For the example in figure 3, the recognition vocabulary would need to contain picture frames apart from faces. Though picture frames may be specific to the newspaper photograph domain, the idea is applicable to other domains with a specific set of interesting objects and a larger set of recognizable objects.

We currently assume an unstructured collection of object locators associated with object classes is available. We are designing a hierarchy of object locators that is linked to our semantic lexicon. This lexicon is organized as a comprehensive ontology of object classes. This will enable us to recognize more object classes using the same primitive object locators. We also assume a corresponding set of image processing tools to verify and compare predetermined attributes of instances of these object classes. Domain specific knowledge allows us to make assumptions regarding object representations and frame of reference for spatial reasoning. Given these assumptions, a naive approach would be:

- Constraint Graph Generation (same as before, constraints are between objects of same/different classes)
- All Object Classes Candidate Generation (apply object locators for all object classes mentioned in the caption)
- Consistent Labeling (verification and recognition)

The above scheme is wasteful, since all object classes needn't be identified for the identification of the target vocabulary. It does not exploit locative constraints to reduce the search space and improve the performance of the object locators/candidate generators.

The caption of the image in figure 3 is *"Sharon Bottoms and April Wade hold a picture of Tyler Doustou"*[1]. Assuming picture frames/portraits in our recognition vocabulary, the relevant constraints generated are

```
included-in(Tyler Doustou, PicFrame1);
left-of(Sharon Bottoms, April Wade)
```

Using the naive approach, labeling is impeded by the failure of the face locator to find Tyler's face. The picture frame also cannot be labeled as no candidate frame object satisfies the *included-in* relation with a face candidate.

Constrained Search

To overcome the above limitations, we choose a more complex approach with some semblance to planning. The idea of using partial interpretation results to localize the search for related objects has been used in several vision systems. For example, constrained search is used by Selfridge [1982] to guide the search for building shadows, but with minimal spatial reasoning. In Section 2 we classified constraints as contextual, spatial and characteristic. When the control algorithm employs any of these constraints to disambiguate candidates, we call them *verification* constraints and when it employs them to locate candidates for an object, we term them *locative* constraints. We assume cost and reliability measures for our object locators and attribute verifiers are available.

The control algorithm loops over three stages:

- **Decision Stage** A complex cost/utility function is employed to decide which object class to detect *next*, given the current state of the constraint graph and labelings. The factors taken into account for this function are the tool cost, reliability, number of objects to be found, size of search region, number of constraints between this class and the objects of interest and the level of interest in this particular class of objects. Some measures are class specific (tool cost), others are instance specific (number of objects) and some are dynamic (size of search region).
- **Labeling Stage:** Candidates for objects of chosen group are located using the collateral information available. These objects, candidates and the constraint relations between the objects are inserted in a constraint graph representing the partial interpretation. Constraint satisfaction techniques are applied to uniquely label the objects. Constraints between objects in the new group and between objects in prior groups and the new ones are satisfied to increment the partial interpretation of the image.
- **Propagation Stage:** Any object which has been labeled uniquely in the previous module potentially constrains the search region for those objects involved in a spatial relation with it. The propagation module computes the search region using the labeled object and the type of spatial relation. So far, this *spatial prediction* has been implemented for binary spatial constraints.

The effect of locating objects of a particular class can then be summarized as (i) potential identification of objects already located, (ii) labeling and partial identification of current class objects and (iii) constraining and prioritizing the search for some other objects.

As we have mentioned, an object locator is parameterized by the number of objects of that particular class. Since the actual routines for object-location are not completely accurate, we might obtain too few or too many object candidates. Information from the spatial prediction stage is used at this point to attempt a solution. The object locator routines are called again on the constrained search region with relaxed parameters if there were initially too few candidates. If there were extra candidates, the information is used to order the candidates and the top choices are picked.

Traditionally, planning in vision has been deliberative, i.e. a plan for the entire task is constructed prior to object location based on probabilistic measures[Garvey, 1976; Ballard *et al.*, 1978]. We employ planning *reactively* to choose amongst locators using a-priori knowledge as well as the observable situation to minimize computation and maximize accuracy.

In addition to verifying spatial relations between object candidates, the **spatial reasoning** module performs the task of spatial *prediction*. In the process of interpretation, any object that becomes uniquely labeled can

[1]Obtained with permission from the Buffalo News

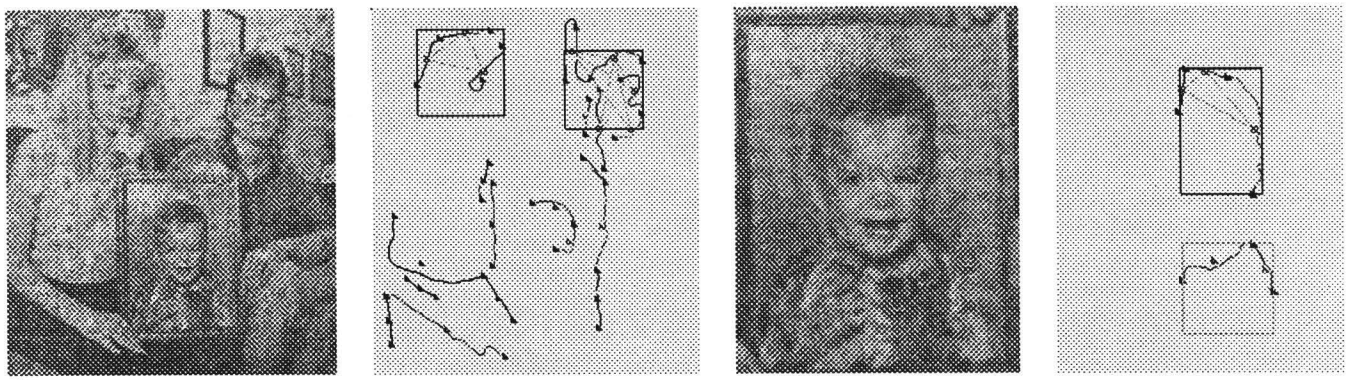

Figure 3: Illustration of constrained search. *From left:* original image and output of face locator for it; constrained search region and output of face locator for it.

serve as a potential *reference* object. Spatial relations between this reference object and other objects *not yet located* are used to generate search areas for those objects. The simplifying assumption of a known fixed reference frame makes this task relatively simple.

As an example of how picture-specific constraints can enable a constrained search for objects, consider Figure 3. The object schema for "picture frames" yields the information that it contains a frame/border and an interior and that the border is typically a rectangle or an oval. The other object schema is for faces. Based on this, the control module invokes the following operations:

- Since rectangles and ovals are easier to find than faces, the decision stage chooses to look for picture-frame objects first. It finds one candidate, a working constraint graph containing a single node is created and labeling is successful. Next, the search region for Tyler's face is constrained because of the relation *included-in*. Control loops back to the decision stage where the next best class (the only one left in this example) is chosen.
- The object locator for faces is now called to find 3 faces. The schema contains information that one face must be in a specified constrained area. Fig. 3 (2nd image) shows the output of the face locator. Since the number of candidates (two) is less than the known number of faces, the locator must be called again. Since neither of the candidates found is in the constrained region, the missing face must be in it, and the locator is called on this sub-image. The last image on the top of the page shows the output. The rectangle in black is the candidate returned by the face-locator and the grey rectangle is a weaker choice not returned in the first pass. The third face candidate is thus located.
- The constraint graph is now augmented with 3 more nodes (representing the 3 persons) and initialized with the new labels and 2 constraints (ref. section3.4) resulting in a successful labeling.

4 Future Work

The cost function evaluated for different classes in the *decision* stage is currently based on weights computed by trial and error. The encoding of relative tool costs and measures should be made more declarative or learned from a training set. We also need to relax our assumption of a known fixed frame of reference and incorporate progress made in the field of spatial representations. The constrained search paradigm is currently being applied only for spatial relations between objects. We need to explore its application to other visual relations, for example, if darker-than(A,B) and A is uniquely labeled, then that constrains the value of the characteristic constraint (brightness) of B, and could be used for a locative purpose.

There are several sources of collateral information which can accompany an image. The task of *co-referencing* words/phrases in the text with their corresponding image areas is greatly facilitated if deictic input is also available. Our system does not currently make use of deictic input.

We are experimenting with several techniques to find faces given poor edge data. These techniques include the use of color, a more sensitive edge detector, and finding facial features (e.g., eyes) directly without locating the face contour. We have recently started experimenting with color images. Color is useful in several ways; it can help the face locator detect boundaries of objects, and allows us to assert and process additional characteristic constraints of objects. An example of such a characteristic constraint for the face object class is hair color.

5 Summary

We have presented a computational model for incorporating picture-specific constraints in the task of image interpretation. We have applied the constraint satisfaction paradigm to this problem. In particular, we have illustrated the use of constrained search in improving the efficiency and accuracy of object locators. A face identification system, *PICTION* has been described, which uses collateral information derived from captions to identify human faces in newspaper photographs. We have identified several limitations of the system and have suggested approaches to overcome them.

Acknowledgments

The authors would like to thank Debra Burhans, Hans Chalupsky, Venu Govindaraju and Mahesh Venkatraman for their input in writing this paper.

References

[Ballard *et al.*, 1978] Dana Ballard, Cristopher Brown, and J. Feldman. An approach to knowledge directed image analysis. In *Computer Vision Systems*, pages 271–282. Academic Press, 1978.

[Burhans *et al.*, 1995] Debra Burhans, Rajiv Chopra, and Rohini Srihari. Domain specific understanding of spatial expressions. In *Proceeding of the IJCAI-95 Workshop on Representation and Processing of Spatial Expressions*, Montreal, August 1995.

[Burns *et al.*, 1984] J. Burns, A. Hanson, and E. Riseman. Extracting straight lines. Technical Report 84-29, University of Massachusetts, Amherst, 1984.

[Chellapa and Kashyap, 1992] R Chellapa and R.L. Kashyap. Image understanding. In Stuart C. Shapiro, editor, *Encyclopedia of Artificial Intelligence.* John Wiley, 2nd edition, 1992.

[Chopra *et al.*, 1995] Rajiv Chopra, Rohini Srihari, and Anthony Ralston. Hyperarc consistency and expensive constraints. Technical Report 95-21, SUNY at Buffalo, 1995.

[Firschein and Fischler, 1987] O. Firschein and M.A. Fischler. *Readings in Computer Vision, Issues, Problems, Principles, and Problems.* Morgan-Kaufmann, 1987.

[Garvey, 1976] J. D. Garvey. Perceptual strategies for purposive vision. Technical Note 117, AI Center, SRI International, 1976.

[Govindaraju *et al.*, 1992] Venu Govindaraju, Sargur N. Srihari, and David B. Sher. A Computational Model for Face Location based on Cognitive Principles. In *Proceedings of the 10th National Conference on Artificial Intelligence (AAAI-92)*, pages 350–355, 1992. San Jose, CA.

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

[Mohr and Henderson, 1986] R. Mohr and Thomas C. Henderson. Arc and path consistency revisited. *Artificial Intelligence*, 28:225–233, 1986.

[Mohr and Masini, 1988] R. Mohr and G. Masini. Good old discrete relaxation. In *Proceedings of ECAI-88*, pages 651–656, 1988.

[Mulder *et al.*, 1988] Jan A. Mulder, Alan K. Mackworth, and William S. Havens. Knowledge structuring and constraint satisfaction: The mapsee approach. *IEEE PAMI-10*, 10(6):866–879, 1988.

[Niemann *et al.*, 1990] H. Niemann, G. F. Sagerer, S. Schroder, and F. Kummert. Ernest: A Semantic Network System for Pattern Understanding. *Pattern Analysis and Machine Intelligence*, 12(9):883–905, 1990.

[Selfridge, 1982] Peter G. Selfridge. *Reasoning About Success and Failure in Aerial Image Understanding.* PhD thesis, U. of Rochester, 1982.

[Srihari and Burhans, 1994] Rohini K. Srihari and Debra T. Burhans. Visual Semantics: Extracting Visual Information from Text Accompanying Pictures. In *Proceedings of AAAI-94*, pages 793–798, 1994. Seattle, WA.

[Srihari *et al.*, 1994] Rohini K. Srihari, Rajiv Chopra, Debra Burhans, Mahesh Venkataraman, and Venugopal Govindaraju. Use of Collateral Text in Image Interpretation. In *Proceedings of the ARPA Workshop on Image Understanding*, pages 897–905, Monterey, CA, 1994.

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

[Strat, 1992] Thomas M. Strat. *Natural Object Recognition.* Springer-Verlag, 1992.