Artificial Intelligence Review **8**: 349–369, 1994–5.

Computational Models for Integrating Linguistic and Visual Information: A Survey

ROHINI K. SRIHARI

Center of Excellence for Document Analysis and Recognition (CEDAR), and Department of Computer Science, State University of New York at Buffalo, UB Commons, 520 Lee Entrance – Suite 202, Buffalo, NY 14228-2567, U.S.A.; E-mail: rohini@cs.buffalo.edu

Abstract. This paper surveys research in developing computational models for integrating linguistic and visual information. It begins with a discussion of systems which have been actually implemented and continues with computationally motivated theories of human cognition. Since existing research spans several disciplines (e.g., natural language understanding, computer vision, knowledge representation), as well as several application areas, an important contribution of this paper is to categorize existing research based on inputs and objectives. Finally, some key issues related to integrating information from two such diverse sources are outlined and related to existing research. Throughout, the key issue addressed is the correspondence problem, namely how to associate visual events with words and vice versa.

Key words: natural language understanding, computer vision, diagram understanding, spatial reasoning, multimedia

1. INTRODUCTION

Much has been said about the necessity of linking language and vision in order for a system to exhibit intelligent behaviour (Winograd 1973, Waltz 1981). A complete natural-language understanding system should be able to understand references to the visual world, especially if it is engaged in discourse or conversation or even reading narratives. Without the ability to visualise, a discourse-understanding system does not have access to a major source of information that speakers may refer to, explicitly or implicitly. Thus, full understanding may not be possible. The same can be said about single-reader situations such as captioned photographs where the ability to 'see' the photograph is crucial in understanding the overall scenario and may in fact be useful in clarifying otherwise ambiguous text. Thus, the integration of language and vision is of great relevance to the task of natural-language understanding.

Integrating language and vision also has implications for knowledge-based vision, since linguistic input (in the form of text or speech) accompanying pictures can be used to dynamically construct scene descriptions. Such a situation could arise in robotics, where a robot is being guided through a visual field by a human who is viewing the same scene on a monitor. These scene descriptions

can then be used by an image-processing system to guide the interpretation of the associated picture.

To date there has been little activity in developing computational models for integrating language and vision. Computer vision has traditionally been viewed as one of the most difficult AI problems; the very modest successes of vision research over the years is a testament to this. The perceived complexity of integrating two intrinsically difficult sub-disciplines, natural language understanding and vision, has kept researchers away from this area. In actuality, the integration of information from these two diverse sources can often simplify the individual tasks (as in collateral text based vision and resolving ambiguous sentences through the use of visual input).

Due to the advent of multimedia processing, there has been an increased focus in this area. There are several applications which can immediately benefit from this new technology. These range from natural language assisted graphics to information retrieval from integrated text/picture databases. Several of these are discussed in this paper.

Whatever be the motivation of the research, the central issue is the *correspondence problem*, namely, how to correlate visual information with words. It is not a simple matter of correlating pictures with words (e.g., nouns); it is necessary to associate visual information with events, phrases or entire sentences thus making the indexing problem very difficult.

Figure 1 depicts the various components of a computational model for integrating linguistic and pictorial information. Depending on the task being attempted, various processing paths may be followed. Each of these tasks involves a mapping of information from a given modality (e.g., text, image, line-drawing) into the appropriate representation in another modality. In cases where multimodal input is present, the task involves consolidating information into a single, unified representation. An example of the latter is the co-referencing task which will be discussed in later sections. What is common to all these tasks is the need for knowledge bases consisting of language models and visual models. These models, combined with domain specific knowledge, enable the mapping of information from one modality into another. Language models may consist of lexicons, grammars and statistical models of language. Visual knowledge consists of the information required to generate and/or recognize objects from line-drawings, raster images or moving image sequences (e.g., video). This could include object schemas (for vision), descriptions of graphical primitives (such as lines, curves and icons) as well as the processing modules associated with these representations.

We organize this survey into three major sections. The first section examines computational models for integrating linguistic and visual information and spans a variety of tasks. This is followed by a computationally motivated discussion of the human cognitive system, one which successfully integrates perceptual information (from the various senses) as well as linguistic information. Finally, the major research issues which arise in the task of integrating visual and linguistic information are discussed.

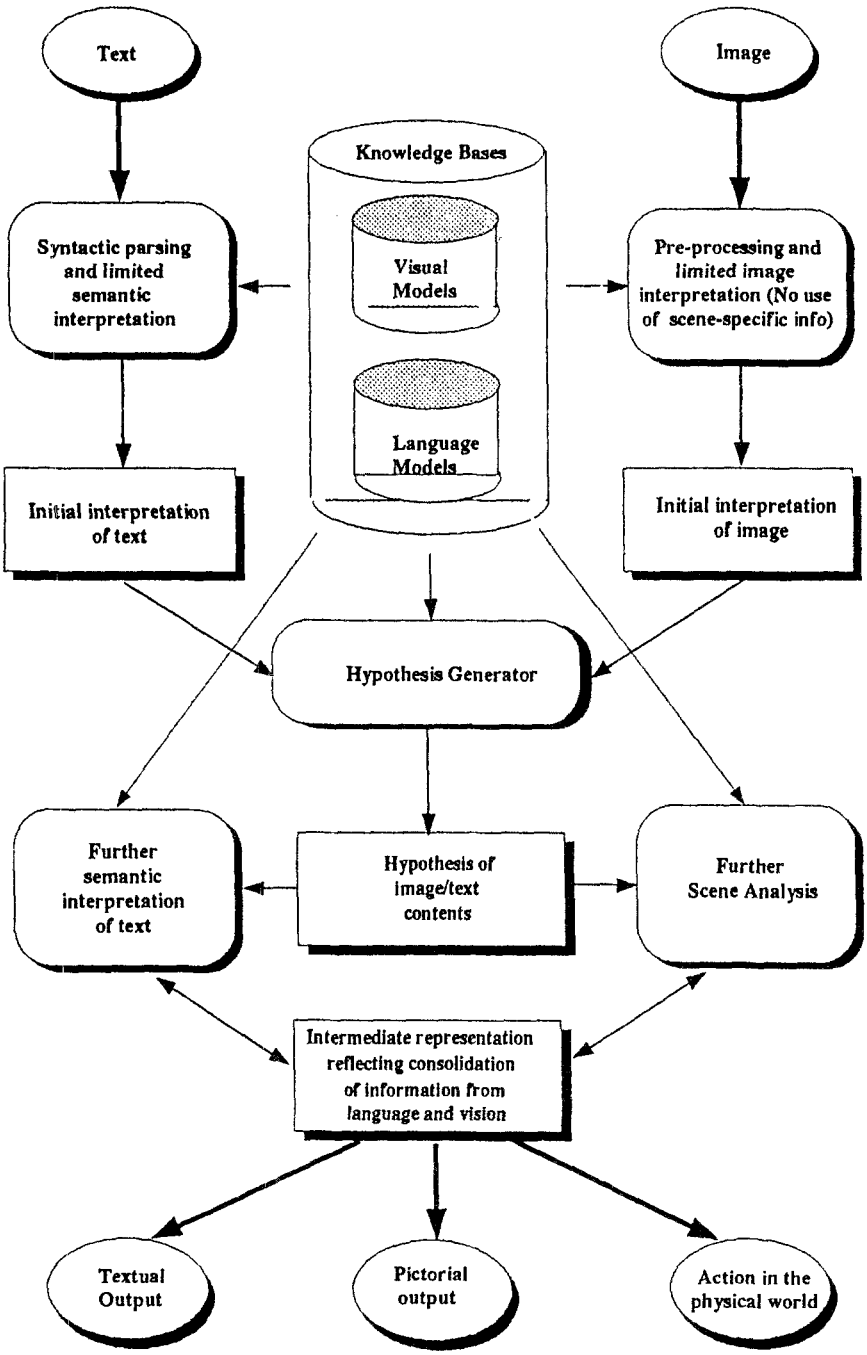


Fig. 1. Computational model for integrating linguistic and pictorial information. Ovals indicate input/output; rounded boxes indicate computational steps; regular boxes indicate derived data.

2. COMPUTATIONAL MODELS

We have classified existing work into two broad areas based on the input types used by these systems as well as their functionality. They are (i) systems that accept either language or visual input but not both, and (ii) systems that deal with both linguistic and pictorial inputs.

Before presenting research in these areas, we mention two systems which are noteworthy since they represent the historical background of this area from two different perspectives. Kirsch (1964) is significant, since he pointed out that in many cases (such as diagrams), text and pictures need to be considered as a whole when it comes to understanding. He described the need for a computer system which could deal with both types of input using uniform methods. The paper is now best known for introducing the idea of syntactic pattern recognition.

Winograd (1973) describes one of the first systems to attempt an integration of language and vision, by accepting block-manipulation instructions in English and displaying the results visually. The ultimate goal was to develop a system for natural-language conversation which could incorporate visual information about the physical world and actually make changes to this world. The system did not have a true vision component however, and relied on symbolic descriptions of objects comprising the physical world. Changes to the world resulted in changes to the configuration of these symbolic descriptions.

2.1. *Language or Visual Input*

The following sections deal with research where either language or visual inputs are used, but not both. However, they rely on integrated visual/language knowledge bases in performing the given task.

2.1.1. *Natural Language Assisted Graphics*

Natural language assisted graphics has been the topic of several recent papers. In such systems, a natural-language sentence is parsed and semantically interpreted, resulting in a picture depicting the information in the sentence. It should be noted that there may be several pictures (in fact, an infinite number) that can be associated with a single sentence. The objective is to produce the most representative picture without generating unintended detail. The key issue that researchers face is understanding spatial language as is illustrated in Figure 2. Understanding this sentence involves correctly interpreting the prepositional phrase 'on the ladder'. The latter implies that the man is on a rung of the ladder and that the ladder is supported by some surface.

In the case of Waltz (1981), the ultimate goal was the ability to reason about the plausibility of utterances. Waltz argued that the ability to visualise both static scenes and events helps us in the understanding process. He proposed a computer model to do this, namely 'event simulations'. An event simulation uses vast amounts of world knowledge and qualitative reasoning in an attempt to visualise (i.e., construct a picture of) a scene based on natural-language input.

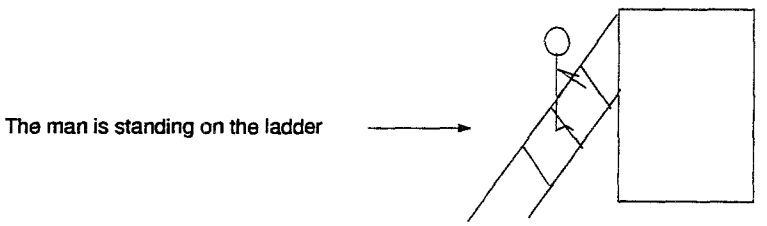


Fig. 2. Illustration of natural language assisted graphics. It is important to infer that the ladder must rest on some surface.

Examples were provided which took simple phrases as input and resulted in line-drawings as output. This was one of the first research endeavours to generate pictures corresponding to an entire sentence.

Where Waltz's work described in general terms a scheme to handle all types of physical events, Adorni *et al.* (1984) describes a specific method of visualising spatial events. The examples in this system were sentences containing prepositional phrases, such as 'the wine is on the table'. Qualitative reasoning suggests that the wine must be in a container (such as a bottle) which is in fact on the table. More recently, Olivier *et al.* (1994) present WIP (Words Into Pictures), a system which automatically generates depictions of spatial descriptions. The focus of this work is on capturing the inherently fuzzy meaning of spatial language. The authors present a method which combines quantitative models as well as a new qualitative model (potential fields) in deriving the semantics of spatial language.

Geller and Shapiro (1987) discuss machine drafting of circuit boards based on natural-language input. The authors describe 'Graphical Deep Knowledge', which is declarative knowledge that can be used for both display and reasoning purposes.

2.2. *Natural-Language Descriptions of Pictorial Information*

Figure 3 illustrates the task of generating a natural-language description of results obtained from a vision system. The problem is to generate a coherent text describing relevant objects, relationships between objects and events which are implicit in the output of the vision system. McDonald and Conklin (1981) describes a system which takes as input a high-level description of the output of a vision system; i.e., a scene is described in terms of the objects present and 3D spatial relationships between them. The objective of the system is to generate coherent text which describes this output. Visual salience is used as a key heuristic in deciding the order of mention of objects. The focus here is entirely on the natural-language generation process, and the visual information is used only as interesting data for this process.

Maddox and Pustejovsky (1987) and Neumann and Novak (1983) describe both systems which deal with time-varying data. They address the problem of recognising events based on intermediate-level visual percepts obtained at

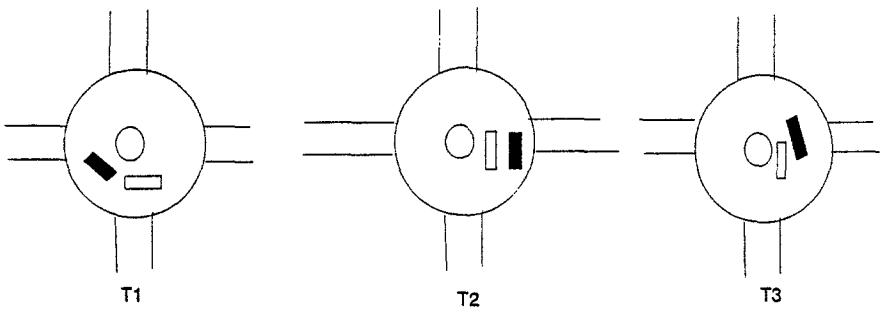


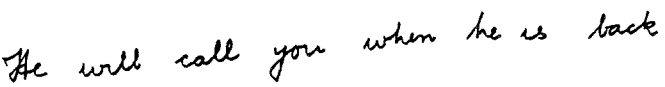
Fig. 3. Generating a natural language description of visual data comprising of moving images; assume that the image sequence was obtained from a camera viewing traffic at a roundabout.

different instances of time. The former uses simulated data and focuses on 'dynamic learning', whereby the system is able to construct or refine event schemata based on an interactive critic's comments. The latter is part of a comprehensive system which uses the output of a motion sequence analyser. The focus is on generating a continuous commentary about events taking place at a busy traffic intersection, which simulates what a human observer might report.

More recently, Howarth and Buxton (1993) discuss the task of selective attention in dynamic vision. The application domain here is once again road traffic. The task is to take as input a temporally ordered stream of 3D pose updates representing traffic movement and (i) identify areas/objects to focus on as well as (ii) determine relationships such as overtaking, and following. The authors conjecture that by limiting the focus of attention, the computational load of interpretation is reduced since it is no longer necessary to consider the interactions of all objects.

2.2.1. *Optical Character Recognition (OCR)*

We have more or less assumed that linguistic input is in the form of text (already separated into words and sentences). In the case of machine-printed and handwritten OCR, the task is to convert spatial data (word images) into a symbolic form, namely, ASCII text. This task also involves the integration of linguistic and visual information. The word images constitute visual input; linguistic information is in the form of statistical models of language which are employed in order to improve the performance of a word recognizer. Khoubyari and Hull (1993) discusses a system which combines visual word-spotting techniques with statistical language models in order to recognize keywords in machine-printed text. Srihari and Baltus (1993) discuss the use of statistical language models in recognizing handwritten text. In Figure 4, it is possible to use syntactic knowledge such as part-of-speech transition probabilities in order to recover the correct sentence from the available word choices.



he with call pen when he us back
she will will you were be is bank

Fig. 4. Digitized image of sentence 'He will call you when he is back' along with the top two word choices for each word image.

2.3. *Incorporating both Linguistic and Pictorial Inputs*

These systems are categorized by their use of both pictorial information as well as language as inputs. They can be classified into four distinct areas: (i) diagram understanding, (ii) map understanding, (iii) computer vision systems, and (iv) multimedia systems. Many of these are within the scope of *document image understanding*. Document image understanding is the task of making a computer understand messages conveyed by printed documents such as newspapers and journals.

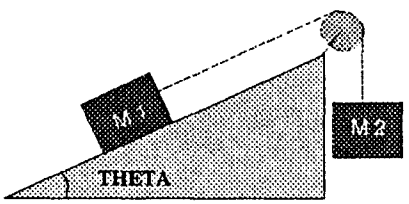
2.3.1. *Diagram Understanding*

A simplistic view of diagram understanding is the conversion of a raster representation to a vector representation: i.e., to convert a binary pixel representation of line-work into a connected set of segments and nodes. Segments are typically primitives such as straight lines, parametric curves, domain-specific graphical icons and text. In addition, (i) portions of the drawing containing text must be converted to ASCII and (ii) graphical icons must be recognized and converted to their symbolic representation. Line segments have parameters such as start positions, extent, orientation, line width, pattern etc. associated with them. Similar features are associated with parametric curves. The connections between segments represent logical (typically, spatial) relationships.

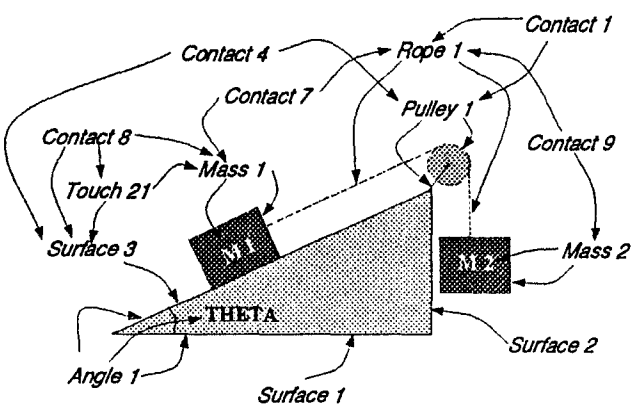
A deeper level of understanding can be attained if groups of primitives (lines, curves, text, icons) are combined to produce an integrated meaning. Consider a dotted line appearing between two words representing city names in a map. If the legend block associates a dotted line with a two-lane highway, one should infer that a two-lane highway exists between the two cities. It is possible to define meanings for documents such as maps, weather maps, engineering drawings, flowcharts, etc. The definition of meaning is somewhat ambiguous in diagrams such as those found in a physics textbook.

Montalvo (1985) introduces the notion of *diagrammatic conversations* between a user and a system, which allows both parties to communicate about different aspects of a diagram using a common language. In order to achieve this, a mapping between symbolic descriptions and visual properties is necessary. The paper discusses how a rich set of visual primitives may be discovered. Examples are presented in the domain of business graphics.

In many diagram understanding applications, the task is to determine the pictorial referent of the entity being described in the text, in other words, the task of *co-referencing*. In Novak and Bulko (1990), the pictorial input consists of symbolic descriptions of lines, circles and rectangles, which, when combined, depict a diagram involving pulleys and ropes. The authors point out that an image pre-processor could easily be trained to recognize basic geometrical shapes. Accompanying the diagram (see Figure 5) is some English text giving further information about the diagram (e.g., the value for theta, an angle depicted in the diagram). A picture parser is able to infer the presence of higher-level entities in the picture such as angles, pulleys, and ropes. This research is interesting, since the parsing of the picture sets up expectations for the parsing of the text. For example, the detection of an angle in the picture may set up an expectation for the value of the angle to be specified in the text. Thus, in some sense, the picture parsing guides the parsing of the text.



Test Problem



Interpretation after Diagram Parsing

Fig. 5. Example illustrating the use of text in diagram understanding. The text accompanying this figure reads: 'Two masses are connected by a light string as shown in the figure. The incline and peg are smooth. Find the acceleration of the masses and the tension in the string for theta = 30 degrees and m1 = m2 = 5 KG.' Example taken from Novak and Bulko (1990).

Rajagopalan (1994) also discusses the co-referencing task with respect to diagrammatic reasoning. The focus of the work is on integrating qualitative spatial and dynamic reasoning for physical systems.

2.3.2. *Map Understanding*

Yokota *et al.* (1984) illustrates the importance of a common intermediate representation in an application which synthesises information from textual weather reports and accompanying weather charts. The system demonstrates its consolidation of the information by presenting visual information verbally and linguistic information pictorially. The latter is illustrated by the generation of a weather chart based on natural-language input from the user. In the case of weather charts, it is possible to derive a straightforward correlation between words and pictures, since every weather concept has a corresponding pictorial symbol.

Reiter and Mackworth (1987) use the domain of geographic maps to illustrate the need for a formal framework for interpretation. The paper defines axioms relating to image-domain knowledge, scene-domain knowledge, and the depiction mapping between the image and scene domains. The authors take an existing map-understanding program, Mapsee, and show how it can be formally defined using these axioms. The claim is that such a formal specification can not only be used in other domains, but is necessary in order to ensure correct performance. The emphasis in this research is on formalising correspondences (once they are known), rather than deriving new ones.

2.3.3. *Computer Vision Systems*

The systems described in this section consider situations where pictures are accompanied by some descriptive text. They illustrate the use of language in constraining the image interpretation task or the use of visual input in constraining the language understanding task.

Abe *et al.* (1981) implement a story-understanding system in which both visual and natural-language input are used to describe the plot. Although the visual processing is relatively simple (line-drawings of a few objects), the system is noteworthy since it attempts to use language to constrain image interpretation *and* visual input to eliminate ambiguities arising in understanding language. Information from language is used initially to guide the search for objects. An example is shown in which an ambiguity in the text is clarified by the detection of a certain object in the picture, thus illustrating the flow of control from the picture to the text.

Truve and Richards (1987) describe an attempt to bridge language and vision through an alternate method of modelling objects. They describe a scheme for transforming image-based descriptions of objects (3D skeleton) into language-based descriptions and vice-versa. Two intemediate representations are described, namely an Object Description Language (ODL) and Connection Tables. ODL is a convenient and formal method of describing how objects are created by the connection of different parts and has the advantage that it is easily translated into English. Thus, ODL serves as an intermediate stage between language and high-level symbolic descriptions of objects. At a lower level, Connection

Tables are an intermediate stage between the ODL representation and the 3D skeleton derived from the image itself. Algorithms are presented in order to convert ODL representations into Connection Tables, thus completing the link between language and vision. The highlight of this work is the establishment of a strong, formal (as opposed to traditionally ad-hoc) method of modelling objects. However, their ideas for transforming language-based descriptions to image-based descriptions extend only to single objects (manufactured, as opposed to natural) rather than a collection of objects (as in a picture).

Srihari (1991), Srihari and Burhans (1994), Srihari (1994) present a computational model whereby textual captions are used as collateral information in the interpretation of the corresponding photographs. The final understanding of the picture and caption reflects a consolidation of the information obtained from each of the two sources and can thus be used in intelligent information retrieval tasks. Although the concept of using collateral information in scene understanding has been explored in systems that use general scene context in the task of object identification, the work described here extends this notion by incorporating picture specific information. A multi-stage system *PICTION* which uses captions to identify humans in an accompanying photograph is described. The author states that this provides a computationally less expensive alternative to traditional methods of face recognition since it does not require a pre-stored database of face models for all people to be identified. A key component of the system is the utilisation of spatial and characteristic constraints (derived from the caption) in labelling face candidates (generated by a face locator). The principal contributions of this work are (i) a theory of systematically extracting visual information from text and representing it as a set of visual constraints, and (ii) an efficient top-down method whereby these constraints can be exploited by an image understanding module engaged in the task of scene interpretation. Step (i) involves the dynamic generation of object and scene *schema*, and employs visual information associated with words in a semantic lexicon.

Finally, Zernik and Vivier (1988) describe a system which, when given scenes of military installations, is able to locate individual objects in the scene (e.g., a tank), based on English sentences containing identifying information (*directive semantics*). An example of the latter is 'The tank is between the hangar and the fueling-depot'. This system uses line-drawings of scenes, rather than digitised images of actual scenes thus simplifying the task of image interpretation.

2.3.4. *Multimedia systems*

Multimedia systems (see Maybury 1993) are those which integrate data from various media (e.g., paper, electronic, audio, video) as well as various modalities (e.g., text, tables, diagrams, photographs) in order to present information more effectively to a user. This is illustrated in Figure 6.

Kobsa *et al.* (1986), Neal *et al.* (1988) and Moore and Swartout (1990) discuss work in the area of intelligent user interfaces. Kobsa *et al.* (1986) describe a system which accepts as input both deictic gestures (i.e., pointing) in reference to a tax form as well as English sentences, and returns as output the exact field in the tax form which the user pointed to. An example input sequence would

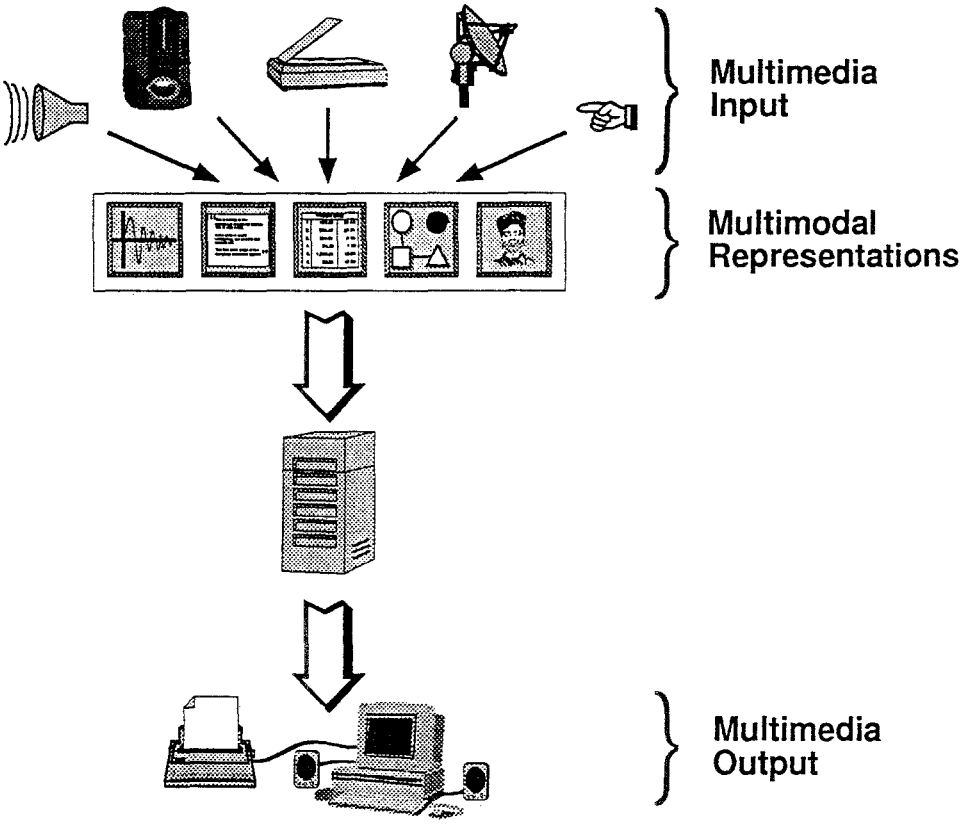


Fig. 6. Multimedia systems deal with different media (paper, audio, video) encompassing several modalities of information (text, diagrams, images).

be the sentence 'Can I add my ACL dues to these membership fees?' combined with the user touching the area of the screen representing some field in the tax form. The challenge is to determine the referent of the pronoun 'these', that is, to determine the field being pointed to. The type of deixis being used in this situation, known as *demonstratio ad oculus*, is distinguished by the fact that the objects on display are visually observable (i.e., have already been introduced) and that the user and the system share a common visual field. This eliminates the need for image interpretation systems. The challenge is to determine which of these pre-identified objects is being referred to, based on simple spatial information (i.e., place on screen which was touched) as well as the linguistic information. Neal *et al.* (1988) describe a system which integrates speech, natural-language text, graphics, and pointing gestures for a human-computer dialogue in the domain of military tactical air control. Moore and Swartout (1990) also address the problem of determining the correct referent but extend the work to generating explanations.

Feiner and McKeown (1991) describe a system in which text and graphics are to be jointly employed in generating helpful explanations for tasks involving

equipment maintenance and repair. For example, the task of clearing a display is explained through a picture of the device in question with the 'CLR' button highlighted (and indicated by an arrow), along with the sentence 'Press the CLR button'. It is assumed that some other process has already generated the content of the explanation and, furthermore, that both the text generator and the graphics generator use the same input. An interesting aspect of this project is the idea of annotating the explanations. If the graphics generator decides to make a knob red (rather than highlighting it), it will record this information. This subsequently causes the text generator to output the phrase 'the red knob'. Thus, we have the notion of bi-directional interaction, an important element of integrating linguistic and pictorial information.

Finally, there has been some recent work in using photograph captions for the purpose of information retrieval. Specifically, Rowe and Guglielmo (1993) discuss a system whereby photographs taken at a naval airbase are entered into a database along with their corresponding captions. Natural language understanding techniques are employed to extract information from the caption which can be later used in retrieving photographs corresponding to user's queries. This avoids the manual step of annotating each photograph such that the relevant 'keys' have the appropriate values. Furthermore, more intelligent querying is enabled by considering the deeper semantics of a caption as opposed to matching against a set of keys. Given the need to convert large collections of photographs to digital format, this technique could be used to catalogue photographs automatically. It should be noted that in cases where captions do not already exist, a user could be asked to input some relevant text. This is more natural than asking users to fill in a form requesting values for all keys.

3. COGNITIVE MODELS LINKING HUMAN PERCEPTION AND LANGUAGE

Since this is a survey of computational models for integrating linguistic and visual information, it is useful to examine theories of how the most efficient computational device, namely the human mind, performs this task. In fact, it is the ease with which humans perform this task that makes it intriguing to attempt a computational model for the same task. Although the word 'perception' could involve any of the five senses as input, for the purpose of the discussion here, it is assumed to be visual perception. There are several viewpoints with regard to this topic and consequently an abundance of literature, but only a few have been discussed here since they attempt to address the correspondence problem form a computational point of view rather than philosophically. We begin with mental imagery since it provides some precedence for attempting the task of integrating visual and linguistic information. This is followed by a discussion of cognitive theories of language understanding and generation.

3.1. *Mental Imagery*

Mental imagery is defined as the human ability to visualise (or construct mental pictures) of various concepts, where the concept can be a simple object (e.g., a

dog) or as complex as an entire sentence. Pinker says that 'unlike (object) recognition, direct inputs and outputs of the imagery system are not known beforehand and must be discovered' (Pinker 1984 p. 37). Although there is agreement among philosophers and cognitive scientists regarding the existence of mental imagery, controversy remains regarding the mechanisms in the brain which support this function.

Kosslyn (1990), gives a convincing argument for the existence of visual imagery and suggests the mechanism by which it operates. He proposes two classes of abilities which mental imagery possesses. The first of these is the ability to use images to retrieve information about the visible (and incidental rather than important) properties of objects (e.g., does a donkey have pointed ears?). The second is the ability to use visual imagery in the course of thinking or reasoning (e.g., imagining how a garden will look after the flowers bloom).

Kosslyn proposes the existence of special hardware in the human brain dedicated to mental imagery, including a special medium in short-term memory (STM) for displaying mental images (known as a visual buffer). This visual buffer has a spatial or array-like structure on which shapes are displayed. He suggests that the long-term memory (LTM) representation of shapes and surface properties used in imagery are the same as those used by our perceptual system in object recognition. More specifically, LTM contains propositional information about the appearance of an object (e.g., a table has four legs) as well as information describing the literal appearance of that object. The former is an object-centred representation, whereas the latter is viewer-centred information. Finally, Kosslyn describes a set of processing modules which operate on the information in STM, LTM, and the visual buffer. Of special interest is the processing module, which can 'produce new combinations of previously viewed objects, including those evoked by verbal descriptions of novel scenes' Kosslyn *et al.* (1984 p. 201).

The work on mental imagery provides some cognitive basis for researchers working in natural language assisted graphics as well as collateral text-based vision.

3.2. *Language Understanding and Generation*

Miller and Johnson-Laird (1976) approach the correspondence problem from a psycholinguistic perspective and attempt to give algorithmic-like descriptions of processes whenever possible. They assume that the perceptual system can make judgments about concepts such as objects, space, time, change, and causation. The output of the perceptual system is input to the conceptual system which attempts to link language and perception.

The authors outline a conceptual theory which is based on (but not limited to) a procedural theory of meaning rather than a verification theory. In the latter, verification is viewed as the fundamental source of evidence for determining linguistic meaning. They state that the conceptual system consists of two parts, a translator and an executor. The translator takes a word or a sentence and produces the procedures which are subsequently executed. These instructions may necessitate a search through perceptual, short-term, or long-term memory. An

augmented transition network is presented as a model for the behaviour of the translator. Some of the instructions include *find* (to search long-term memory), *store* (to affect long-term memory), *generate* (for visual imagery), and *identify* (which references both long-term memory as well as perceptual memory).

As a first step, the authors examine how objects are related to the words that name them. They adopt the view that a label is associated with a perceptual 'paradigm' rather than with an object or class. A perceptual paradigm consists of information which is necessary in order to identify an object (this is known as an 'object model' in computer vision). Additionally, there is also functional information associated with a label. These two types of information are combined together into a 'schema' which incorporates different sets of perceptual paradigms. This type of information is represented in long-term memory.

In summary, the authors recognise that integration is performed at a common intermediate level, namely, the conceptual level. Second, the representation of lexical concepts makes it plain that both functional information as well as information relating to perception must be associated with a word. This could be taken as a guideline for constructing integrated language/vision lexicons.

Jackendoff (1987) describes a theory for correlating words and images on a lower level. In particular, he outlines a method whereby an intermediate representation (referred to as 'conceptual semantics') provides the link between language and Marr's computational theory of vision (Marr 1982). He suggests that the human ability to categorise objects and recognise individuals is due to the conceptual primitives 'TOKEN' and 'TYPE', where the first is used to label an individual object and the second is used to recognise and label categories of objects. In order to recognise objects, Jackendoff suggests that object descriptions similar to Marr's 3D models are associated with classes of objects, thus providing a link between the perceptual system and the conceptual level. Furthermore, both visual imagery and object recognition share the same 3D representation of objects.

He suggests an extension to Marr's 3D representation of objects which would include path information (helpful in describing the motion of objects) and would extend object-internal coordinate axes to the space exterior to an object. The latter would prove helpful in spatial reasoning. The ideas he expresses are powerful, since his method would establish a correlation between language and pictures at the conceptual level. Implementation of these ideas, however, requires a very sophisticated vision system which can disambiguate among models at very detailed levels of description.

Jackendoff deals with the correspondence problem mainly at the single-word level (both nouns and verbs) but does not extend the discussion to establishing a correspondence between a sentence/phrase and the complex scene which it may evoke.

4. RESEARCH ISSUES IN INTEGRATING LINGUISTIC AND VISUAL INFORMATION

There are several areas in which further research is necessary if truly integrated systems are to be realized. Some of these have already been encountered earlier but are elaborated on here.

4.1. *Integrated Language/Vision Knowledge Bases*

The development of large scale linguistic knowledge bases has been the focus of recent research. Wordnet (see Beckwith *et al.* 1991) is a large ontology of words based on psycholinguistic principles and is frequently used in natural language applications. Unfortunately, such a large scale ontology of visual object descriptions has not been investigated although preliminary work is reported in Srihari and Burhans (1994). If available, a visual ontology would permit (i) objects to be detected at various resolutions, (ii) objects sharing several visual properties to be grouped, and (iii) new objects to be classified based on their visual properties (i.e., visual learning).

It is not sufficient to develop independent ontologies corresponding to language and vision. In order for language understanding systems and vision systems to interact, the knowledge bases they use must be integrated. Integrated knowledge bases when combined with world knowledge as well as spatial and qualitative reasoning modules, would facilitate:

- Language understanding systems to access visual knowledge thereby allowing the generation of visual analogues corresponding to sentences or phrases. The latter could be used by diagrammatic reasoning systems.
- Vision systems to access linguistic knowledge thereby allowing natural language descriptions of visual data.
- Enforcing uniform word meanings across language processing and vision systems.

Integrated language/vision knowledge bases would also facilitate automatic learning of new linguistic concepts (i.e., words) and new visual concepts (i.e., object schemas). Siskind (1990) reports on preliminary attempts to design a system which learns the meaning of new words. It takes as input both linguistic input, expressed as a set of sentences, and visual input, expressed as a sequence of conceptual structures describing visual scenes. The system produces a lexicon as output which reflects its 'learning' of unknown words. It would be interesting to see if such a system could be extended to use line-drawings or raster images as its visual input rather than higher-level conceptual structures.

A major issue is the modelling of dynamic events. If one is to describe action in the world, or moving sequences of images, it is necessary to represent dynamic events such as running, overtaking, and jumping. This requires the matching of low-level dynamic visual percepts with intermediate concepts in event frames.

Although some work has been reported in creating integrated knowledge bases for certain specialized domains (e.g., weather maps, traffic scenes), in general this remains an open problem.

4.2. *Visual Semantics: Extracting and Interpreting Visual Information in Language*

One of the key problems arising in tasks such as co-referencing, collateral text-based vision and language assisted graphics is deriving visual semantics from the text. Visual semantics refers to information present in textual input which is useful in understanding an accompanying picture or in generating a visual analogue. Deriving visual semantics involves lexical, syntactic and semantic processing of text. A significant portion of visual semantics involves the interpretation of spatial prepositions, a topic which has been actively investigated and discussed later in this paper. However, there has been little work in the spatial interpretation of open class words (e.g., wearing, holding) which convey a wealth of visual information.

4.3. *Mapping Visual Data Into Symbolic Representations*

Computer vision systems deal with the problem of image interpretation, that is, deriving the meaning of a scene in terms of the objects present and their interrelationships. Although such a high-level symbolic representation is the ultimate goal for picture processing systems, achieving it is extremely difficult due to the noise present in the input as well as the enormous size of the search space. For many applications however, it is sufficient to derive more modest symbolic representations of visual data. Examples of this include colour histograms, texture measures as well as skeletons (see Chang 1989). Nakatani and Itoh (1994) report on an image retrieval system based on colour indexing where users communicate their queries in English. Words such as 'darker' are interpreted based on the colour model being employed. Significant work is being conducted in deriving such representations by the image database research community.

4.4. *Spatial Reasoning*

The semantic interpretation of prepositional phrases constitutes a formalism for establishing correspondence between language and pictures and is the focus of much research. It is discussed extensively in Herskovits (1986) and Talmy (1983) from the point of view of understanding language.

An intelligent spatial reasoning module must allow for arbitrary shape representations. Many have expressed the viewpoint that it is difficult to formulate a general purpose qualitative spatial model since it would require knowledge of exact shapes of objects. However, recent work in qualitative spatial reasoning has focused on approximating shapes of objects. Abella and Kender (1993) present a method for 'fuzzifying' qualitative spatial prepositions such as 'near' and 'along'. The proposed method works for irregular shaped objects and uses a shape representation based on the center and elongation axes of the object. Rajagopalan (1994) discusses a model for integrating qualitative spatial and dynamic reasoning about physical systems; it extends current methods by allowing for effects of translational or rotational motion on the spatial state. Generalising these models sufficiently to enable their use in a natural language understanding system still remains a challenge.

4.5. *Control structures for Integrating Language and Vision*

In tasks involving back and forth processing both language and visual data, sophisticated control structures are required which are able to exploit constraints obtained from one modality in the processing of the other. This may entail redesign of existing NLP and vision systems such that:

- the system is modularized thereby permitting an external control mechanism to call on low-level processes;
- lower-level processes can exploit contextual (top-down) constraints by instantiation of appropriate parameters and other mechanisms;
- costs and weights are associated with individual processes thereby allowing an external control mechanism to evaluate the various control choices at any given time;
- processing can be suspended and reactivated when useful information is obtained.

General purpose control models such as constraint satisfaction may be appropriate for such tasks. However, issues such as adding and retracting constraints, utility and costs associated with satisfying constraints and effectively integrating top-down and bottom-up control must all be explored. More importantly, it should be feasible to derive the required constraints from the input and exploit them efficiently at the appropriate time. Other control structures such as blackboard systems have previously been used in applications (such as speech recognition) where it is necessary to integrate information from diverse sources.

4.6. *Knowledge Representation*

One of the issues to be considered is whether the intermediate representation is flexible enough to handle both visual and linguistic information. This applies primarily to those systems that employ both pictorial and linguistic input.

Allen in Allen (1987) mentions three types of knowledge which are used in natural-language processing, namely syntactic knowledge, word-sense knowledge, and world knowledge. Syntactic knowledge refers to the permissible structures of sentences. Word-sense knowledge (which is a type of semantic knowledge) refers to the meanings of words (in given situations) and the associations between words. When combined, the above three types of knowledge permit a representation of the initial meaning of the sentence to be constructed.

The most commonly used knowledge representation (KR) formalisms for representing semantic knowledge are logic and semantic networks (Sowa 1991). Logic programming languages such as Prolog have been successfully employed in natural language understanding systems. SNePS (see Shapiro and Rapaport 1990) is a fully intensional, propositional semantic-network processing system in which every node represents a unique concept. Especially noteworthy is the natural-language parsing and generating facility which is part of SNePS (Shapiro 1982). A vital component of a KR scheme is the inference system that operates on the knowledge base. Rule-based systems such as Prolog employ a knowledge base of axioms and rules and use resolution or natural deduction to carry out reasoning. The majority of existing semantic networks use path-based infer-

ence (or 'inheritance'), where the presence of a specified path of arcs is used to infer facts. SNePS, on the other hand, allows both rule-based inference (called 'node-based inference') as well as path-based inference.

Knowledge representation in vision refers to the process of modelling physical constraints of the real world such that this information can be used by a computer vision system in 'understanding' a scene. In Ballard and Brown (1982), the authors specify several criteria which are essential to a knowledge base to be used in computer vision: (i) represent analogical, propositional, and procedural structures, (ii) allow quick access to information, (iii) be easily and gracefully extensible, (iv) support inquiries to the analogical structures, (v) associate and convert between structures, and (vi) support belief maintenance, inference, and planning.

One of the earliest knowledge-representation schemes proposed for computer vision was the idea of 'frames' Minsky (1975). More recently, *schemas*, which are based on frames, have been used widely in the vision community. A schema is 'a collection of information about an object, including the relations between parts of the object, a description of the geometric structure of the object and strategies for recognition of the object' (see Weymouth 1986). Information such as size, colour, and sub-components is typically declarative information, whereas the description of an object's shape is often expressed procedurally. There is also the need for a schema hierarchy, where components of objects may be object schemas themselves. Finally, schema networks are collections of schemas capturing relations among individual objects and expected contents.

Semantic networks such as KL-ONE have also been successfully used in computer vision systems to represent schemas. LOOM (see MacGregor 1991) is a semantic network system which has been employed in natural language applications such as machine translation, computer vision applications, and applications requiring both natural language processing as well as visual processing (see Srihari 1994). LOOM is a high level programming language and environment for constructing knowledge based systems which is based on the KL/1 family of semantic networks. It provides an object-oriented model specification language, classification based inference and multiple programming paradigms in a framework of query based assertion and retrieval.

We have discussed the need for a rich knowledge representation formalism which could effectively model knowledge from either modality as well as represent the consolidated knowledge. Such representations are required if the objective is to build systems which reason with both language and visual data. However, such powerful representation schemes have considerable overhead associated with them and thus may not be practical (or required) for use in applications such as information retrieval. It may be necessary to condense the information and use data models suited for database systems. The next section discusses such types of representations.

4.7. *Consolidating Linguistic and Visual Information*

In order to enhance the presentation of information to a user, multimedia systems require information pertaining to a query to be presented in various modalities

and media. Significant research has focused on (i) linking semantically related information from various modalities, (ii) developing methods for effectively presenting the information to a user and (iii) developing techniques for automatically deriving the links in (i). We have already discussed how systems which perform co-referencing as well as collateral based vision systems assist in task (iii) above. There are many issues which remain pertaining to how the information should be consolidated to permit efficient access. Multimedia data modelling (see Ishikawa *et al.* 1993) continues to be an actively researched area.

5. SUMMARY

In this paper, we have presented research on computational models for integrating language and vision. Both implemented systems as well as computationally motivated research in human cognition have been presented. Some of the key issues in integrating knowledge from such diverse sources have been outlined and related to existing research.

Although the ultimate goal of developing an intelligent agent which has both language and perceptual abilities remains elusive, progress has been made towards integrated language/vision systems in restricted domains and tasks. The advent of multimedia processing has given some impetus for continuing research in this field since several applications can immediately benefit from this technology. In that respect, there has been recent interest in developing systems which correlate visual data with linguistic data at a very high-level (e.g., using natural language to describe picture attributes in a text/image retrieval system). In order to develop truly intelligent systems however, the difficult problem of relating language and visual abilities at the basic levels of cognition must be investigated. For the latter, insight offered by cognitive scientists may prove to be useful.

The nature of this research is such that it spans many subfields of artificial intelligence, including natural-language processing, computer vision, spatial reasoning and knowledge representation and inference. Based on a survey of existing research, it is apparent that a new area of research is emerging, one which attempts to link the above subfields in a cohesive manner.

REFERENCES

Abella A. & Kender R. (1993). Qualitatively Describing Objects Using Spatial Prepositions. In Proceedings of *The Eleventh National Conference on Artificial Intelligent (AAAI-93)*, 536–540. Washington, DC.

Allen, J. (1987). *Natural Language Understanding.* Benjamin/Cummings: Menlo Park, CA.

Adorni, G., Di Manzo, M. & Giunchiglia, F. (1984). Natural Language Driven Image Generation. In Proceedings of *COLING*, 495–500.

Abe, N., Soga, I. & Tsuji, S. (1981). A Plot Understanding System on Reference to Both Image and Language. In Proceedings of *IJCAI-81*, 77–84.

Ballard, D. H. & Brown, C. *Computer Vision.* Prentice Hall: New Jersey.

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

Chang, Shi-Kuo (1989) *Principles of Pictorial Information Systems Design*. Prentice-Hall.

Feiner, S. K. & McKeown, K. R. (1991). Automating the Generation of Coordinated Multimedia Explanations. *IEEE Computer* **24**(10): 33–41.

Geller, J. & Shapiro, C. (1987). Graphical Deep Knowledge for Intelligent Machine Drafting. In Proceedings of *The Tenth International Joint Conference on Artificial Intelligence (IJCAI-87)*, 545–551. Morgan Kaufmann: Los Angeles, CA.

Hearth, R. & Burton H. (1993). Selective Attention in Dynamic Vision. In Proceedings of *The 13th International Joint Conference on Artificial Intelligence (IJCAI-93)*, 1579–1584.

Herskovits, A. (1986) *Language and Spatial Cognition*. Cambridge University Press.

Ishikawa, H., Suzuke, F., Kozakura, F., Makinouchi, A., Miyagishima, M., Izumida, Y., Aoshima, M. & Yamane, Y. (1993). The Model, Language, and Implementation of an Object-Oriented Multimedia Knowledge Base Management System. *ACM Transactions on Database Systems*. **18**(1): 1–50.

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

Kosslyn, S. M., Brunn, J., Cave, K. & Wallach, R. (1984). Mental Imagery Ability. In Pinker, S. (ed.) *Visual Cognition*, 1–63. MIT Press: Cambridge Mass.

Khoubyari, S. & Hull, J. (1993). Keyword Location in Noisy Document Images. In Proceedings of *The Second Annual Symposium on Document Analysis and Information Retrieval*, 217–231.

Kirsch, R. A. (1964). Computer Interpretation of English Text and Picture Patterns. *IEEE Transactions on Electronic Computers* **13**: 363–376.

Kobsa, A. *et al.* (1986). Combining Deictic Gestures and Natural Language for Referent Identification. In Proceedings of *COLING*, 356–361.

Kosslyn, S. M. (1990). Mental Imagery. In Osherson, D. A. *et al.* (eds.), *Visual Cognition and Action*, 73–97. MIT Press: Cambridge Mass.

MacGregor, R. (1991). The Evolving Technology of Classification-Based Knowledge Representation Systems. In *Principles of Semantic Netwerks: Exploration in the Representation of Knowledge*, 385–400. Morgan Kaufmann: Los Angeles, CA.

Marr, D. (1982). *Vision*. W. H. Freeman: San Francisco.

Maybury, T. (ed.). (1993). *Intelligent MultiMedia Interfaces*. AAAI Press/The MIT Press.

McDonald, D. & Conklin, E. J. (1981). Salience as a Simplifying Metaphor for Natural Language Generation. In Proceedings of *AAAI-81*, 49–51.

Minsky, M. (1975). A Framework for Representing Knowledge. In Winston, P. H. (ed.), *The Psychology of Computer Vision*, 211–277. McGraw-Hill Book Company: New York, NY.

Miller, G. A. & Johnson-Laird, P. N. (1976). *Language and Perception*. The Belknap Press of Harvard University Press: Cambridge, MA.

Montalvo, S. F. (1985). Diagram Understanding: The Intersection of Computer Vision and Graphics. A.I. Memo 873, Massachusetts Institute of Technology.

Maddox, A. B. & Pustejovsky, J. (1987). Linguistic Descriptions of Visual Event Perceptions. In Proceedings of *The Ninth Annual Cognitive Science Society Conference*, 442–454, Seattle.

Moore, J. D. & Swartout, W. R. (1990). Pointing: A Way Toward Explanation Dialogue. In Proceedings of *The Eighth National Conference on Artificial Intelligence (AAAI-90)*, 457–464.

Novak, G. S. & Bulko, W. C. (1990). Understanding Natural Language with Diagrams. In Proceedings of *The Eighth National Conference on Artificial Intelligence (AAAI-90)*, 465–470, Boston.

Neal, J. G., Dobes, Z., Bettinger, K. E. & Byoun J. S. (1988). Multi-Modal References in Human-Computer Dialogue. In Proceedings of *AAAI-88*, 819–823. Morgan Kaufmann.

Nakatani, H. & Itoh, Y. (1994). An Image Retrieval System that Accepts Natural Language. In *Working Notes of the AAAI-94 Workship on Integration of Natural Language and Vision Processing*, 7–13.

Neumann, B. & Nova, H. (1983). Event Models for Recognition and Natural Language Description of Events in Real-World Image Sequences. In Proceedings of *IJCAI 1983*, 724–726.

Olivier, P., Maeda, T. & Tsujii, J. ichi (1994). Automatic Depiction of Spatial Descriptions. In Proceedings of *AAAI-94*, 1405–1410. Seattle, WA.

Pinker, S. (ed.). (1984). *Visual Cognition*. MIT Press: Cambridge Mass.

Rajagopalan, R. (1994). A Model for Integrated Qualitative Spatial and Dynamic Reasoning about Physical System. In Proceedings of *AAAI-94*, 1411–1417. Seattle, WA.

Rowe, N. & Gugliemo, E. (1993). Exploiting Captions in Retrieval of Multimedia Data. *Information Processing and Management* **29**(4): 453–461.

Reiter, R. & Mackworth, A. K. (1987). A Logical Framework for Depiction and Image Interpretation. Technical Report 88-17. The University of British Columbia.

Shapiro, S. C. (1982). Generalized Augmented Transition Network Grammars for Generation from Semantic Networks. *The American Journal for Computational Linguistics* **8**(2): 12–25.

Shapiro, S. C. & Rapaport, W. J. (1990). The SNePS Family. CS Technical Report 90-21, SUNY at Buffalo.

Siskind, J. M. (1990). Acquiring Core Meanings of Words, Represented as Jackendoff-Style Conceptual Structures, from Correlated Streams of Linguistic and Non-Linguistic Input. In Proceedings of *The 28th Annual Meeting of the Association for Computational Linguistics*, 143–156.

Sowa, J. F. (1991). *Principles of Semantic Networks: Exploration in the Representation of Knowledge*. Morgan Kaufmann: Los Angeles, CA.

Srihari, R. K. (1991). PICTION: A System that Uses Captions to Label Human Faces in Newspaper Photographs. In Proceedings of *The 9th National Conference on Artificial Intelligence (AAAI-91)*, 80–85. Anaheim, CA.

Srihari, R. K. & Baltus, M. (1993). Incorporating Syntactic Constraints in Recognizing Handwritten Sentences. In Proceedings of *The International Joint Conference on Artificial Intelligence (IJCAI-93)*, 1262–1267.

Srihari, R. K. & Burhans, D. T. (1994). Visual Semantics: Extracting Visual Information from Text Accompanying Pictures. In Proceedings of *AAAI-94*, 793–798. Seattle, WA.

Srihari, R. K. (1994). Use of Collateral Text in Understanding Photos. *Artificial Intelligence Review (special issue on integration of NLP and Vision)*, (this volume).

Talmy, L. (1983). How Language Structures Space. In Pick, H. & Acreolo, L. (eds.) *Spatial Orientation: Theory, Research, and Application*, 225–282. Plenum: New York.

Truve, S. & Richards, W. (1987). From Waltz to Winston (via the Connection Table). In Proceedings of *The First International conference on Computer Vision*, 393–404. Computer Society Press.

Waltz, D. L. (1981) Generating and Understanding Scene Descriptions. In Webber, Bonnie & Sag, Ivan (eds.) *Elements of Discourse Understanding*, 266–282. Cambridge University Press: New York, NY.

Weymouth, T. E. (1986). Using Object Descriptions in a Schema Network for Machine Vision. PhD thesis, University of Masschussetts at Amherst.

Winograd, T. (1973). A Procedural Model of Language Understanding. In *Computer Models of thought and Language*, 152–186. W. H. Freeman and Company: San Francisco.

Yokota, M., Taniguchi, R. & Kawaguchi, E. (1984). Language-Picture Question-Answering Through Common Semantic Representation and its Application to the World of Weather Report. In Bolc, Leonard (ed.) *Natural Language Communication with Pictorial Information Systems*. Springer-Verlag.

Zernik, U. & Vivier, B. J. (1988), How Near Is Too Far? Talking about Visual Images. In Proceedings of *The Tenth Annual Conference of the Cognitive Science Society*, 202–208. Lawrence Erlbaum Associates.