CHAPTER 18

# KNOWLEDGE-BASED MULTIMEDIA SYSTEMS[1]

**Jeannette G. Neal**
*Calspan Corporation*

**Stuart C. Shapiro**
*State University of New York at Buffalo*

Multimedia systems hold the promise of great benefits in terms of increasec productivity, efficiency, effectiveness, and information enjoyment. Multi- media systems promise to provide the needed increase in the bandwidth o information exchange between humans and computers, and to enhanc human understanding of complex information through better presentatior technologies and appropriate combinations of these technologies for infor mation presentation. However, before these promises can be fulfilled, ther are many problems that need to be solved. Many of these problems are i technology areas such as multimedia document authoring, multimedi information and document storage and management, search techniques computer-supported collaborative work, and multimedia human-compute interaction. The field of artificial intelligence will help provide solutions t these problems.

1 This research was supported in part by the Defense Advanced Research Project Agency and monitored by the Rome Air Development Center under Contract Nc F30603-87-C-0136.

This chapter discusses the possible role of artificial intelligence in multimedia systems. A concept for a multimedia system is presented that provides an integrated work environment with a human-computer interface designed as an intelligent agent with the ability to communicate and make presentations in coordinated multiple media/modalities. The objective is to integrate the various subsystems and functionality that a user needs in a workstation environment, to simplify operator interaction with sophisticated computer systems, and to minimize the time and effort spent by the user on manipulating the interface. A working prototype system, called CUBRICON (CUBRC [2] Intelligent CONversationalist), is also discussed.

This chapter also reviews some of the current research being conducted in the area of artificial intelligence applied to multimedia systems. Future directions are also presented.

## 18.1 INTRODUCTION

Multimedia systems hold the promise of great benefits such as increasing peoples' productivity, efficiency and effectiveness, and increasing the utility and enjoyment of our vast information resources. Multimedia systems promise to provide the needed increase in the bandwidth of information exchange between humans and computers, and to enhance human understanding of complex information through better presentation technologies and appropriate combinations of these technologies for information presentation. The extent to which these promises are fulfilled depends on continued improvement in hardware technology, development of much needed supportive software technology, and the growth of a community of trained multimedia authors and technologists.

The scope of today's multimedia systems is very limited, and the functionality of the different types of systems is not integrated to form a productive workplace. In fact, multimedia means different things to different people. To some it means video for conferencing, to others it means hypermedia documents, and to others it means multimedia human-computer dialogue. Also, some people view multimedia documents as static and fixed, and others view documents and data as "live." For example, Clark states that multimedia should be referred to as " 'interactive electronic presentation (IEP)' to describe a collusion of sounds and images elicited from a piece of (electromechanical) machinery by the user's persistent activity" [1, p. 75]. However, Clark's definition includes only the concept of self-contained multimedia "books" to be consumed by "readers" or "viewers," and he states that an IEP is closed and finite. On the other hand, Boy [2,3] and Cornell, Suthers, and Woolf [4] stress that documents and data should be treated as "live" or dynamic.

Certainly the view of documents and data being static and fixed is inadequate for people engaged in productive activity or problem-solving tasks. People will need to be able to locate, retrieve, use, save, and possibly manipulate relevant multimedia documents/data in an environment that will support the accomplishment of their tasks, possibly in cooperation with others.

We take the position that multimedia does not simply mean self-contained static documents, nor does it simply mean that computer-based video is used to provide a "media space" to support cooperative work among co-located or remotely located people [5,6].

Our concept of a multimedia system is that of an integrated work environment with a human-computer interface designed as an intelligent agent with the ability to communicate and make presentations in coordinated multiple media/modalities. The objective is to integrate the various subsystems and functionality that a user needs in a workstation environment, to simplify operator interaction with sophisticated computer systems, and to minimize the time and effort spent by the user on manipulating the interface. The human-computer interface should have the ability to: conduct dialogue with the user; act as an intelligent assistant for accessing application systems; accept and understand input expressed in multimodal language; decide how information and responses are to be presented to the user, including the selection of media/modalities for information presentation, composition, and presentation of the output in multiple modalities; adhere to respected human factors guidelines for human-computer interaction and information presentation; and support cooperative work with others. This system concept is discussed in Section 18.3.

This chapter discusses the possible role of artificial intelligence in multimedia systems and some of the current research being conducted. We also present our concept of a multimedia system and discuss its architecture and functionality. Although necessary hardware is becoming widely available at reasonable cost, application software and trained personnel are in short supply. These related problems are discussed in Section 18.2. Our system concept and prototype are discussed in Section 18.3. Related research is described in Section 18.4. Conclusions and future directions are presented in Sections 18.5 and 18.6, respectively.

## 18.2 PROBLEMS FACING MULTIMEDIA SYSTEMS

Although great benefits are to be gained from multimedia systems, their incorporation into the workplace, school, and home is not an easy task. The

[2] Calspan—UB Research Center (UB is the State University of New York at Buffalo)

requisite hardware is becoming widely available at reasonable cost, but other problems remain to be solved. These problems are primarily in two areas, personnel and technology:

1. Personnel
   - ☐ There is a lack of people trained in the development and management of distributed databases and document repositories, and
   - ☐ In the words of Grimes and Potel, "a fundamental problem afflicts multimedia authoring—not enough people have the necessary skills" [7, p. 25].
2. Technology
   - ☐ There is a lack of software designed to integrate, control, coordinate, manage, and adapt the various media for human-computer interfaces.
   - ☐ There is a lack of support software for facilitating the authoring, composition, and production of multimedia documents.
   - ☐ There is a lack of support technology in the area of multimedia data and document storage and manipulation.
   - ☐ There is a lack of search and pattern recognition capability for locating information and/or documents that are of interest in multimedia storage facilities.
   - ☐ There is a lack of software support technology for group decision making and cooperative work, especially in the application of multimedia technology to cooperative decision making and work.

All of the tasks listed above are difficult and provide good candidates for ıpplication of artificial intelligence technology to help solve the problem.

## 18.3 THE ANATOMY OF AN INTELLIGENT MULTIMEDIA SYSTEM

ıs mentioned briefly in Section 18.1, our concept of a multimedia system onsists of an integrated work environment with a human-computer interıce designed as an intelligent agent with the ability to conduct dialogue ⁄ith the user in coordinated multiple media/modalities. The human-comıuter interaction is modeled on the manner in which two or more people ıturally communicate in coordinated multiple modalities when working ⁄ith graphics, video, and other devices at hand. The system should have ıe ability to:

- ☐ Conduct dialogue with the user:
  - — Adhere to respected principles of conversation [8], and
  - — Adhere to respected human factors guidelines for human-computer interaction and information presentation, including maintaining the context of the dialogue and maintaining consistency in displays and presentations.
- ☐ Maintain knowledge and belief models to enable the system to understand user inputs and compose system outputs:
  - — Track and model the dynamic focus of the dialogue in order to maintain context during the dialogue.
  - — Model the user's task(s) and the state of the user's accomplishments and progress with respect to the task(s).
- ☐ Maintain knowledge bases of information about:
  - — Modalities and user interaction,
  - — World knowledge, and
  - — Application-specific knowledge.
- ☐ Act as an intelligent assistant for accessing and using application systems, through such activities as:
  - — Assisting the user in finding relevant information on topics of interest.
  - — Assisting the user with finding, selecting, and accessing appropriate tools to apply to the task.
  - — Assisting and guiding the user in the accomplishment of tasks.
  - — Providing explanations and multimedia presentations to aid in user comprehension of relevant information.
- ☐ Accept and understand input expressed in multimedia language.
  - — Provide the user with flexibility in the media that is selected and combined for expressing input to the system.
- ☐ Decide how information and responses are to be presented to the user:
  - — Select modalities/media for information presentation.
  - — Compose the output in multiple modalities.
  - — Present the multimedia output in a coordinated manner.
- ☐ Manage the windows by intelligently performing the window operations (i.e., creation, placement, sizing/resizing, moving, iconization, retrieval, and destruction) to relieve the user of the burden of performing these chores.

### 18.3.1 Intelligent Multimedia System Design

Figure 18.1 provides an overview of our design for an intelligent multimedia system. An implemented system, called CUBRICON, has been developed as a proof-of-concept prototype as part of the Intelligent MultiMedia Inter-

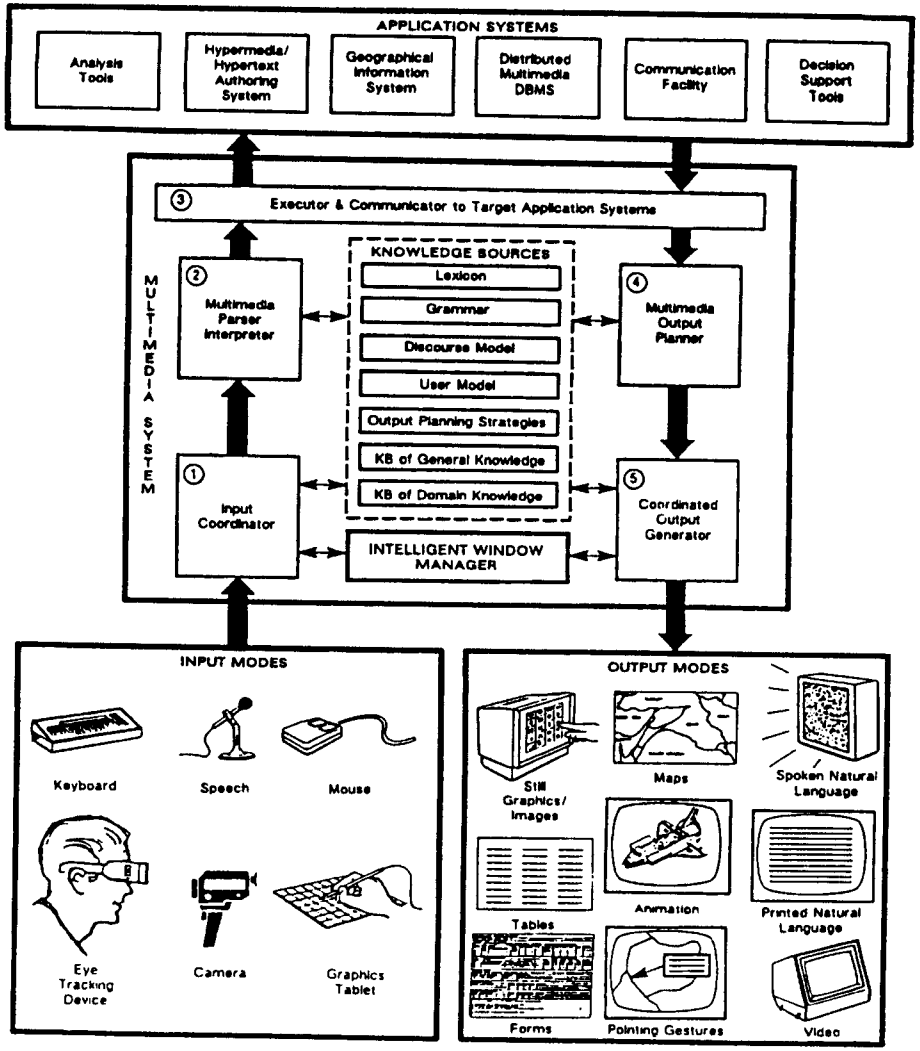


**Figure 18.1** Multimedia system design

faces Project [9,10,11,12,13]. The CUBRICON prototype includes implementation of all the components shown in the box labeled "Multimedia System." The application used for the CUBRICON prototype was that of air force mission planning. The input modes that were implemented in the CUBRICON system were speech, keyboard, and mouse. All of the output modes shown in Figure 18.1 were implemented in the CUBRICON system except for video, and the implemented graphics and animation capability was fairly simple.

The CUBRICON system design is based upon an integrated use of communication modes or media, whether verbal, visual, tactile, or gestural. Human beings primarily communicate with each other via written and spoken natural language and gestures, supplemented with pictures, diagrams, video, and other sounds. The CUBRICON system design provides for the use of a unified multimedia language. Input and output streams are treated as compound streams with components corresponding to different media. This approach is intended to imitate, to a certain extent, the ability of humans to simultaneously accept input from different sensory devices (such as eyes and ears), and to simultaneously produce output in different media (such as voice, pointing motions, and drawings). The CUBRICON system includes: (a) language parsing and generation that processes and supports synchronized multimedia input and output streams, (b) knowledge representation and inferencing to provide reasoning ability, (c) knowledge bases and models to provide a basis for its decision-making ability, and (d) automated knowledge-based medium selection and formulation of responses.

CUBRICON possesses the following critical functionality. CUBRICON:

- ☐ accepts and understands multimedia input such that references to entities in a natural language sentence can be accompanied by coordinated simultaneous pointing to the respective entities on a graphics display:
  - — is able to use a simultaneous pointing reference and natural language reference to disambiguate one another when appropriate;
  - — infers the intended referent of a point gesture which is inconsistent with the accompanying natural language
- ☐ automatically composes and generates relevant output to the user in coordinated multimedia:
  - — automatically selects appropriate output media/modalities for expressing information to the user, with the selection based on the nature of the information, discourse context, and the importance of the information to the user's task;
  - — uses its media/modalities in a highly integrated manner including simulated parallelism;
  - — judges the relevance of information with respect to the discourse context and user task and responds in a context-sensitive manner;
  - — adheres to respected human factors guidelines for human-computer interaction and information presentation; these guidelines include: (i) maintain the context of the user/computer dialogue, (ii) maintain consistency throughout a display, and (iii) maintain consistency across displays.
- ☐ automatically performs the window manipulation operations (i.e., creation, placement, sizing/resizing, moving, iconization, retrieval, and destruction) so as to relieve the user of the need to manipulate the interface.

CUBRICON accepts input from three input devices: speech input device, keyboard, and mouse device pointing to objects on a graphics display. CUBRICON produces output for three output devices: high-resolution color graphics display, monochrome display, and speech output device. The primary path that the input data follows is indicated by the modules that are numbered in Figure 18.1:

1. Input Coordinator,
2. Multimedia Parser Interpreter,
3. Executor/Communicator to Target System,
4. Multimedia Output Planner, and
5. Coordinated Output Generator.

The Input Coordinator module accepts input from the three input devices and fuses the input streams into a single compound stream, maintaining the temporal order of tokens in the original streams. The Multimedia Parser/Interpreter is an augmented transition network (ATN) that has been extended to accept the compound stream produced by the Input Coordinator and produce an interpretation of this compound stream. Appropriate action is then taken by the Executor module. This action may be a command to the mission planning system, a database query, or an action that entails participation of the interface system only. An expression of the results of the action is then planned by the Multimedia Output Planner for communication to the user. The Output Planner is a generalized ATN that produces a multimedia output stream representation with components targeted for different devices (e.g., speech device, color graphics display, monochrome display). This output representation is translated into visual/auditory output by the Coordinated Output Generator module. This module is responsible for producing the multimedia output in a coordinated manner in real time (e.g., the Planner module can specify that a certain icon on the color graphics display must be highlighted when the entity represented by the icon is mentioned in the simultaneous natural language output).

The CUBRICON system includes several knowledge sources to be used during processing. The knowledge sources include:

- ☐ a lexicon,
- ☐ a grammar defining the language used by the system for multimedia input and output,
- ☐ discourse model,
- ☐ user model,
- ☐ a knowledge base of human-computer interaction knowledge, including output planning strategies to govern the composition of multimedia responses to the user,
- ☐ a knowledge base of information about generally shared world knowledge, and
- ☐ a knowledge base of information about the specific task domain of tactical air control.

These knowledge sources are used for both understanding input to the system and planning/generating output from the system. They are discussed in more detail later.

The CUBRICON system is implemented on a Symbolics Lisp Machine with a color graphics monitor, a monochrome monitor, and a mouse pointing device. Speech recognition is handled by a Dragon Systems VoiceScribe 1000. Speech output is produced by a DECtalk speech production system. CUBRICON software is implemented using the SNePS semantic network processing system [14,15,16], an ATN parser/generator [17], and Common Lisp. SNePS is a fully intensional propositional semantic network and has been used for a variety of purposes and applications [16,18,19,20,21]. SNePS provides:

- ☐ a flexible knowledge representation facility in the semantic network formalism;
- ☐ representation of rules in the network in a declarative form so they can be reasoned about like any other data;
- ☐ a bidirectional inference subsystem [22] which focuses attention towards the active processes and cuts down the fan-out of a pure forward or backward chaining;
- ☐ a simulated multiprocessing control structure [23];
- ☐ special nonstandard connectives [24] to model human reasoning processes.

### 18.3.2 The Multimedia

The CUBRICON design and implementation incorporates the following media or modalities: spoken natural language, typed or printed natural language, pointing gestures, geographical maps, color graphic pictorial displays, tables, and "fill in the blank" forms. This list does not exhaust the possibilities, of course, but provides a good variety with which to prove our concept and upon which to build. Other media, such as video and eye-tracking devices, were not used in the prototype but would be a natural extension of the system and are included in Figure 18.1.

One of the significant features of the CUBRICON system is that it not only generates output in multiple modalities, but also decides which modalities to use and how to use and combine them. CUBRICON modality

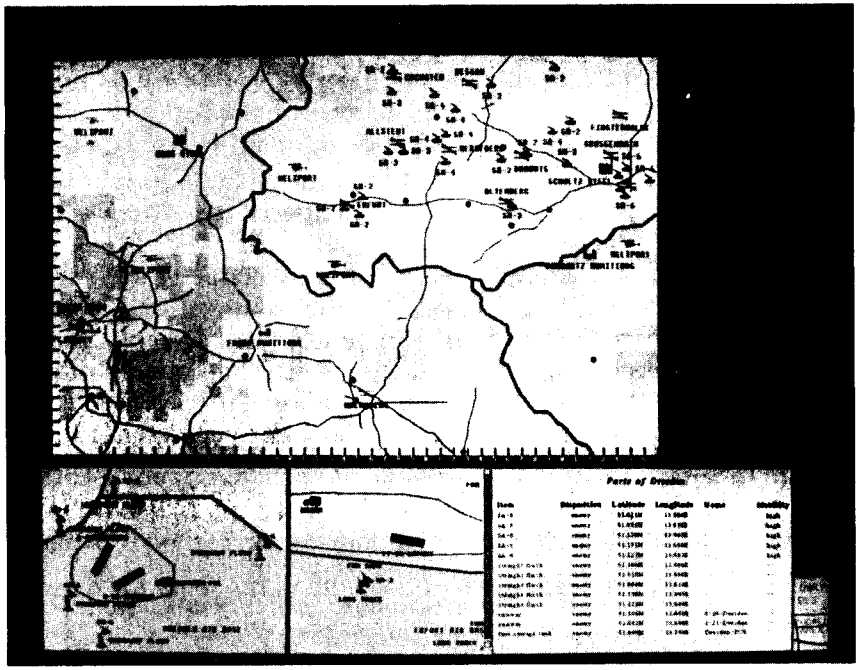

**Figure 18.2** Example CUBRICON displays

selection is primarily based on the nature and characteristics of the information and the purpose for which the modality is being used. Our system design is based on the premise that graphic/pictorial presentation is always desirable. The following is a list of the CUBRICON modalities and a brief summary of the selection criteria.

1. **Color graphics:** Selected whenever the CUBRICON system knows how to represent the information pictorially.
2. **Geographic maps:** Selected when the information is geographically locative or has a locative attribute. An example is shown in Figure 18.2.
3. **Table:** Selected when the values of common attribute(s) of several entities must be expressed.
4. **Forms:** A predefined form is selected when the task engaged in by the user requires the form. An example modeled on one of the forms used by air force mission planners is shown in Figure 18.3.
5. **Animation:** Simple animation is used for information or objects that can be visually presented and which are temporally changing or moving.
6. **Deictic gestures:** Selected for emphasis or to call the user's attention to one or more objects on the screen(s).

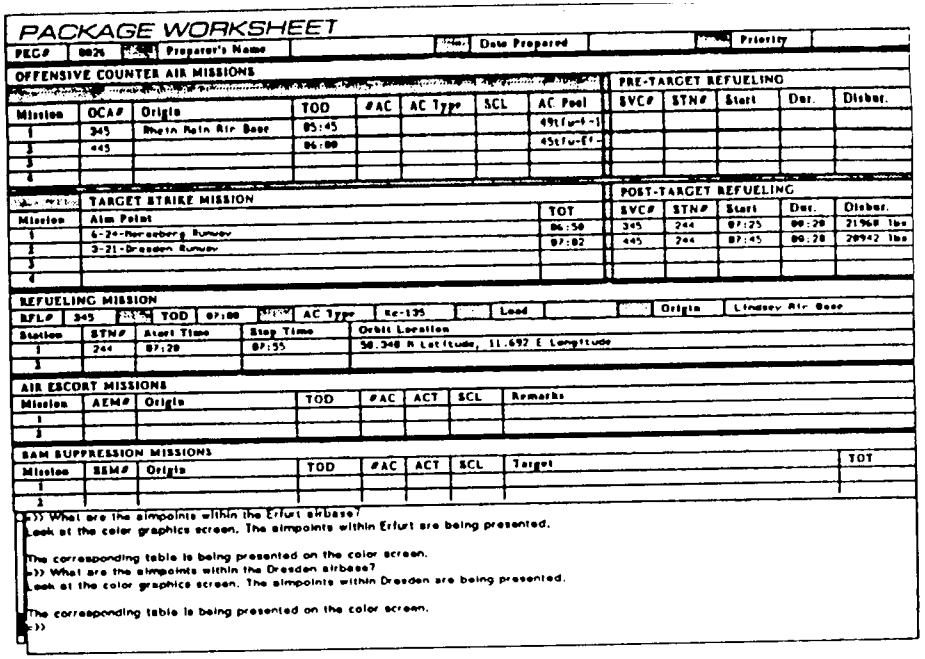


**Figure 18.3** The CUBRICON mission planning form

7. **Natural language:** Selected for the expression of a proposition, relation, event, or combination thereof, when the types of knowledge structures being expressed are heterogeneous. Natural language can be presented in either spoken or written form.

   Printed natural language (printed on the screen) is selected for longer technical responses that would strain the user's short-term memory if speech were used (see [25]).

   Spoken natural language is used in a manner that is designed to avoid overwhelming the user's short-term memory. It is selected for:

   - ☐ Dialogue descriptions to assist the user in comprehending the presented information. These include explanations of graphic displays or display changes and verbal highlighting of objects on the displays (e.g., "The enemy airbases are highlighted in red").
   - ☐ Warnings to alert the user of important events that have taken place or are about to take place (e.g., new critical information comes into the application system database and the system notifies user: "The XXX airbase has been damaged by enemy shellfire").
   - ☐ Informing the user about the system's activity (e.g., "I'm still working" when the user must wait for output from the system).

- ☐ Short expressions of relatively nontechnical information that can be remembered when presented serially (e.g., a "yes"/"no" answer to a user's question).

Most frequently, multiple modalities are desirable to present a body of information to the user. For example, to inform the user about the movements of a certain tank battalion, a desirable presentation might be an explanation delivered in combined spoken speech and coordinated drawing on a graphic map display showing movements of the battalion, as well as a printed textual summary with ancillary information on the monochrome display. The multiple modalities should be selected to complement and enhance one another. Andriole [26] has used graphic equivalence effectively using dual displays or split screens to present the same material in different forms to aid user comprehension and problem-solving performance. We are not restricting the system to presenting the *same* material in different forms, but, instead, our system presents related material or different aspects of a given event or concept in different forms/modalities (as appropriate based on the nature and characteristics of the information). We are also not restricted to graphic display presentations.

### 18.3.3 Knowledge Sources for Multimedia Interaction

The CUBRICON system includes several knowledge sources for use in multimedia language understanding and production. These knowledge sources are a lexicon and grammar; a discourse model; a user model; a knowledge base of human-computer interaction knowledge, including output planning strategies to govern the composition of multimedia responses to the user; a knowledge base of information about generally shared world knowledge; and a knowledge base of information about the application task domain used in this research effort, namely, tactical air control.

#### The Lexicon and Grammar

A lexicon is the collection of all the tokens or signals that carry meaning in a given language. The CUBRICON system's lexicon consists of words, graphic figures, and pointing signals. The grammar defines how the morphemes, tokens, and signals of the lexicon can combine to form legal composite language structures. An example of a multimodal language structure that is legal according to the CUBRICON grammar is a noun phrase. A noun phrase consists of the typical linguistic syntax (e.g., determiner followed by zero or more modifiers followed by a noun) accompanied by zero or more pointing signals (pointing to objects on the graphics display). The lexicon and grammar together define the multimodal language used by the system.

#### The Discourse Model

Continuity and relevance are key factors in discourse. Without these factors, people find discourse disconcerting and unnatural. The attentional discourse focus space representation [27,28,29,30] is a key knowledge structure that supports continuity and relevance in dialogue. CUBRICON tracks the attentional discourse focus space of the dialogue carried out in multimedia language and maintains a representation of the focus space in two structures:

1. a main dialogue focus model, which includes those objects and propositions that have been explicitly expressed (by the user or by CUBRICON) via natural language or by a pointing or direct manipulation gesture, and
2. a display model, which represents all the objects that are "in focus" because they are visible on one of the monitors.

CUBRICON is based on the premise that visual communication is an integral part of language, along with natural language and other forms of text and pointing. The CUBRICON system treats objects presented visually on the graphics displays as having been intentionally "expressed" or "mentioned." All objects on the graphics display are "in focus," and CUBRICON maintains a representation of all these objects in the display model. The display model consists of two levels: 1) a list of the windows on each monitor, and 2) a list of all the objects that are visible in each window. This display model is used in a manner that is analogous to the use of the main dialogue focus model.

When processing the user's input, the dialogue attentional focus space representation is used for determining the interpretation of anaphoric references [29] and definite descriptive references [31] expressed by the user in natural language. In the CUBRICON system, the main dialogue focus model is consulted in determining the referent of a pronoun. In the case of a definite reference, if an appropriate referent is not found in the main dialogue focus model, then CUBRICON consults the visual display model. The motivation for this is the fact that when a person expresses a definite reference such as "the airbase" with just one such object in view (as on a graphics display) and none have been verbally discussed, then the person most likely refers to the one in view even though he or she might know about several others.

The discourse model is used during output generation also. When CUBRICON composes a reference for an object as part of a natural language sentence, for example, it consults the discourse model. If the object is represented in the display model as being visible in one of CUBRICON's windows, then the system uses a deictic dual-media expression to refer to the object in the output sentence. The deictic expression consists of a phras

such as "this airbase" and simultaneously blinks/highlights the airbase as its means of pointing to the object. If the object is the most salient of its gender according to the main focus model, CUBRICON uses a pronoun to refer to the object.

### The User Model

Many aspects of a user are highly relevant to human-computer interaction and user modeling is an active area of research [32,33,34,35,36]. Relevant aspects of the user include his or her level of expertise in the current task, perspective based on his or her role, value system, degree and nature of impairedness due to fatigue or illness, and preferences concerning mode of communication. To address all of these aspects of user modeling is, of course, beyond the scope of this chapter. The aspects of the user that seemed most relevant in our research and which are modeled in the CUBRICON user model are: 1) the degree of importance that the user attaches to the different object types as a function of task, which we call the user's *entity rating system*; and 2) the stage of the current task on which the user is currently engaged.

CUBRICON includes a representation of the user's entity rating system as a function of the task being addressed by the user. For a given task in the process of being carried out by the user, the entity rating system representation includes a numerical importance rating (on a scale from zero to one) assigned to each of the object types used in the application task domain. The numerical rating assigned to a given object type represents the degree of importance of the object to the user. Associated with the entity rating system is a *critical threshold* value: Those objects with a rating above the critical threshold are critical to the current task and those with ratings below the threshold are not. The CUBRICON design provides for the entity rating system representation to change automatically under program control in the following manner: 1) when the user's task changes, the system replaces the current entity rating list with the standard initial rating list for the new task; and 2) when the user mentions an entity whose rating is lower than the critical threshold, then its rating is reset to be equal to the critical threshold to reflect the user's interest in the entity and its seeming relevance to the current task from the perspective of the user. In the current implementation, CUBRICON performs the second function listed above, but the implementation of the first is not complete.

The user's entity rating system plays an important role in composing responses to the user: 1) the entity rating system representation is used in determining what information is relevant in answering questions or responding to commands from the user, 2) the entity rating system is used in selecting ancillary information to enhance or embellish the main concept being expressed and to prevent the user from making false inferences that he might otherwise make, and 3) the entity rating system is also used in organizing the form in which information is presented.

As an example of 1) above, if the user instructs the system to "Display the Fulda Gap Region," CUBRICON uses the entity rating system representation to determine what objects within the region should be displayed. If the user is a military mission planner, then displaying all the country cottages in the region, for example, is irrelevant. The objects to display are those that are relevant to the job of the mission planner. Thus, the objects that the system selects from its database for display are airbases, missile sites, targets, etc.

CUBRICON includes a representation of the current task in which the user is engaged. CUBRICON's mode of response to the user is affected by whether or not the user's task has just changed. The CUBRICON team is developing a task hierarchy: a decomposition of the user's main tasks into subtasks. This a priori task knowledge can be used by CUBRICON to help track the discourse focus, manage the displays, and anticipate the needs of the user.

### Knowledge Bases: General and Application-Specific

The CUBRICON system includes knowledge bases containing general and application-specific information. General information includes world knowledge applicable across different task domains, while application-specific information is applicable to the particular task on which the user is engaged.

Crucial information included in the knowledge bases is information concerning the visual presentation or verbal expression of the objects/concepts known to the system. This information includes the words and symbols used to express an object, which symbols are appropriate under which conditions, and when particular colors are to be used.

An important component of the application-specific knowledge base is a representation of the different types of mission plans that the user would be engaged in constructing. The knowledge base includes a model or structure to represent each type of mission plan (e.g., Offensive Counter Air, Refueling) and the components of each type of plan (e.g., aircraft, airbase, temporal information, flight path), as well as specific instances of plans that have been developed by the user.

## 18.3.4 Multimedia Language Understanding

A user communicates with the CUBRICON system using natural language and gestures (pointing via a mouse device). Typically, the user speaks to the system, but keyboard input is just as acceptable. The use of pointing combined with natural language forms a very efficient means of expressing a definite reference. This enables a person to use a demonstrative pronoun as a determiner in a noun phrase and simultaneously point to an entity on the graphics display to form a succinct reference. Thus, a person would be able to say "this SAM" (surface-to-air missile system) and point to an object on the display to disambiguate which of several SAM systems is meant. The

alternative, using natural language only, would be to say something like "the SAM system at 10.35 degrees longitude and 49.75 degrees of latitude" or "the SAM system just outside of Kleinburg." The use of pointing references combined with natural language is efficient, since the cognitive process of generating the dual-media reference would be much shorter than the generation of the reference using natural language only. The result is a reduction in the cognitive workload for the user.

The CUBRC team has developed a formal grammar defining the syntax of the multimedia language. The grammar is implemented in the form of a generalized ATN. The traditional ATN, which takes a linear textual input stream, has been modified so that it takes a multimedia input stream with components from the different input devices. Input from the devices is accepted and fused into a compound stream, maintaining the information as to which point gesture(s) occurred with (or between) which word(s) of the sentence. Each noun phrase or locative adverbial phrase can consist of zero or more words of text, along with zero or more pointing references to objects on the displays (there must be at least one point reference or one word). The pointing input that is a component of a noun phrase or locative adverbial can occur anywhere within the phrase: as the first token(s), between the natural language words of the phrase, or as the last token(s).

In the CUBRICON system, four types of objects can be referred to via pointing:

- ☐ Geometric points within any window (e.g., a map or graph);
- ☐ Objects represented by icons;
- ☐ Table entries; and
- ☐ Windows on the monitors.

CUBRICON accepts interrogative, imperative, and declarative sentences, although the most commonly used are interrogatives and imperatives. The following are illustrative examples. These inputs presuppose that a map is displayed on the color graphics screen with icons representing various objects. Each "<point>" represents a point to an object or location on one of the graphics displays.

INTERROGATIVE:

"Where is the 43rd Tank Battalion?"

"What is the mobility of this <point> SAM?"

"Is this <point> the base for these Troop Battalions $<point>_1$ $<point>_2$ $<point>_3$?"

IMPERATIVE:

"Display the East-West Germany Region."

"Display the aimpoints within this <point> airbase."

"Present the OCA1001 mission plan."

Use of such dual-media references entails certain problems, in that: 1) a point by the user can be ambiguous if he or she points to a location where two or more graphical figures overlap (one could be part of another as in the case of a hubcap being part of a car wheel) or two or more icons overlap (they could be closely located), and 2) the user can inadvertently miss the object at which he or she intended to point. CUBRICON uses semantically based techniques for resolving such issues. These problems and the CUBRICON solution techniques are discussed in [9,10,12,13].

An important feature of a multimedia system such as CUBRICON is that the user and the system have significant flexibility with respect to their manner of communication. To refer to a particular object such as an airbase, for example, one could use its proper name, refer to it by its location, or simply point to it on a map. Similarly, for intangible objects such as a mission plan: The plan could be described using paragraphs of text or by the use of a form such as shown in Figure 18.3.

Another important feature of the CUBRICON system is that it is a unified system in which various displays and presentations reflect a single integrated underlying reality. This underlying reality is represented in the CUBRICON knowledge base, which is central to the system. So, for example, as a user builds up an air force mission plan, he or she may input decisions verbally as in "Make Nuremberg the origin airbase of the OCA345 mission" or he or she may input such information via the form shown in Figure 18.3. The key central representation of information, such as the elements of the mission plan, is in the knowledge base. This knowledge base is updated when the user makes such inputs. Existing visual representations on the displays are updated also.

An interesting way for the user to input a plan decision such as the one mentioned above is to make a very terse, efficient spoken input with accompanying point gestures such as the following:

USER: "Enter this <point-map-icon> here <point-form-slot>.

This example illustrates that CUBRICON enables the user to use point gestures in conjunction with more than just one phrase of a sentence and that the point gestures may access different types of windows on different monitors. In this example, the user's first point gesture touches an object on a map display on the color graphics CRT (Figure 18.2) and the second selects a slot of the mission planning form on the monochrome CRT (Figure 18.3). One of CUBRICON's features that is critical to its ability to process this input is that its display model contains representations of the objects displayed visually in each of the windows of each CRT. The object representations in the display model are the same representations as in the knowledge base, so that each object has a unique representation and we avoid the problem of tracking and maintaining multiple representations of an object. The knowledge base is shared by all the modules of the CUBRICON system. If the <point-map-icon> selects the Nuremberg airbase on the map of Figure 18.2 and the <point-form-slot> touches the "origin airbase"

slot on the mission planning form of Figure 18.3, CUBRICON builds the knowledge base structure which represents the assertion that Nuremberg is the airbase from which the particular mission will be flown. The visual version of the form seen by the user is also updated.

### 18.3.5 The CUBRICON Intelligent Window Manager

One of the important technologies in human computer interaction is the use of windows to enable the computer to help users manage and access several sources of information on one screen, much as office workers typically organize desk space into separate areas for organizing papers by use category [37].

The CUBRICON Intelligent Window Manager (CIWM) [12,38] was designed to automatically perform all window placement and manipulation functions within the CUBRICON system. The decision to automate window management functions was based on the premise that this would reduce the efforts required of the user for window management, and thus free the user's mental and temporal resources for task domain activities. As the problems and application tasks confronting computer users become more complex and information-intensive, the potential of this approach for improving overall human-system performance is enhanced. Bly and Rosenberg [39] found that, for a database management task, almost half of the user's time is spent in managing the window-based interface. If their findings are representative of all or most computer-based tasks that use windowing systems, the concept of automated window management offers great potential for increasing human-computer effectiveness on these tasks. In addition, automatic window management offers a critical capability for applications in which the user's hands and cognitive resources are not available, such as in the cockpit environment or other situations in which the person is performing simultaneous tasks.

The CIWM is a knowledge-based component that automatically performs window management functions on CUBRICON's color and monochrome screens, including window creation, sizing, placement, removal, and organization. These operations are accomplished by the CIWM without direct human inputs, although the system provides for user override of the CIWM decisions.

Important CIWM features, including window layout, placement, importance, sizing, and window iconization, are discussed briefly in the following paragraphs.

**Window Layout:** The CIWM combines tiled and overlapping layout approaches to form a hybrid window configuration management methodology, allowing CUBRICON to realize the advantages of both types of windowing systems, while minimizing the disadvantages. The CIWM always prefers tiled windows, but allows a window to overlap adjacent windows when necessary based on window contents and the task at hand.

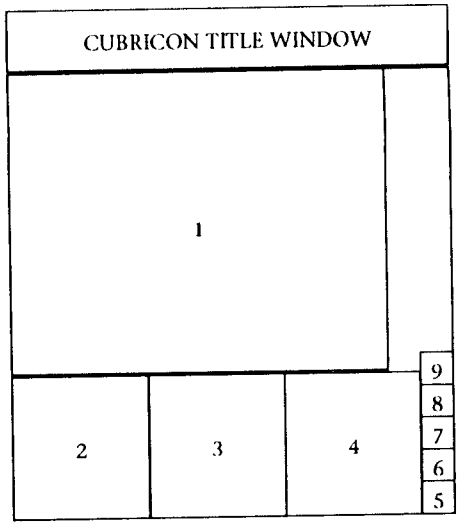


**Figure 18.4** The CUBRICON preferred window positions

**Window Placement:** Window placement combines heuristics for each window type (e.g., prefer to place geographic maps on the color CRT, forms on the monochrome CRT, tables on the monochrome CRT) with placement logic that addresses the problem of placing a window within a given CRT screen. For a given CRT screen, placement of windows is based on window type and on relative importance and relationships among the displayed windows. In placing a window on one of the CRTs, the CIWM checks whether one of the four preferred predefined window positions shown in Figure 18.4 is available. If so, the CIWM puts the new window in the available position. If all the window positions are full, the CIWM removes (and iconizes) the least important window to make room for the new window.

**Window Importance:** The importance of each window is computed by an algorithm which is a function of:

time of creation,
contents,
recency of use,
time since last interaction,
frequency of use, and
context (or relation to the ongoing dialogue).

**Window Sizing:** The size of a map window is determined by a combination of its function and an algorithm that computes the minimum necessary size based on clutter analysis. (This algorithm may dictate a window size that results in the new window overlapping adjacent windows.)

**Window Iconization:** As the number of windows increases, a limited number of the most useful windows are kept open, while those of lesser importance are removed and transformed to labeled icons. Billingsley [40], for example, has discussed the use of icons to remind users of closed but available windows. This maximizes the visibility of available information and maintains an organized display. Animation is used to portray the window-to-icon transformation for the user.

The CIWM is discussed in more detail in [12,38] with interactive examples, evaluation results, and limitations and applicability of the research.

### 18.3.6 Multimedia Presentations for Space/Time-Dependent Activities

Just as it accepts multimedia input, CUBRICON generates multimedia output that combines spoken natural language, visually displayed natural language, graphics, simple animation, and deictic gestures. An important feature of the CUBRICON design is that the output modalities are composed and coordinated by a single generator providing a unified multimedia language with components (e.g., speech and graphics) synchronized in real time. CUBRICON uses a generalized augmented transition network (GATN) [17] for multimodal language generation. This GATN generates multimedia language "on the fly" from knowledge base structures or representations—that is, the presentations are not "canned." The knowledge base representations are based on generic structures and relations with generic "slot" names (e.g., subtype, supertype, subtask, supertask, name, property-name, property) so that the multimodal generation components can be applied and used with other applications. In this section, we focus on the multimedia presentation of space/time-dependent activities. Other aspects of CUBRICON's multimedia generation capability are discussed in [11,12,13].

CUBRICON's output generation component composes and produces multimedia presentations that are designed for the explanation and presentation of activities consisting of events with spatial and temporal attributes. The prototype implementation has been applied to the presentation of OCA (Offensive Counter Air) missions that include flight path traversals with related activities, such as the striking of targets and the airborne refueling of strike aircraft. This multimedia presentation component uses two critical concepts: *task hierarchy* and *granularity measure*. These are discussed in the next paragraphs, followed by discussion of discourse structure and output composition.

#### Task Hierarchy

CUBRICON's multimedia presentation component uses a knowledge base representation of a task hierarchy and the interrelationships and properties of the tasks. This information is used to define the concept of granularity in the CUBRICON system. For the mission planning application domain

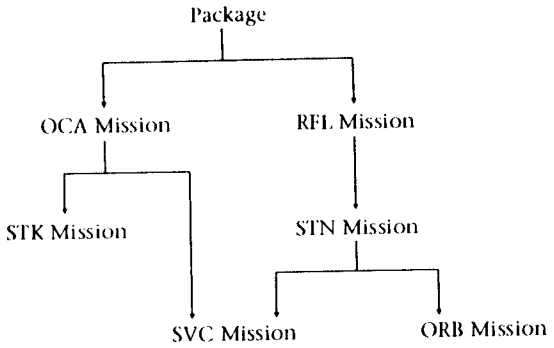


**Figure 18.5** Hierarchy of mission types

used in this project, the tasks represented in the hierarchy in CUBRICON's knowledge base are tactical air force missions. This knowledge base includes information about the different types of missions and their interrelationships and properties. The primary mission types that are modeled in CUBRICON's current knowledge base are shown in the hierarchy of Figure 18.5. The dashed lines in this figure represent the relation of parent to child (or supertask to subtask). The knowledge base also includes information about mission details such as, for an OCA mission, the airbase from which the mission aircraft originate, the aircraft unit flying the mission, the type of aircraft used, the time of departure of the aircraft, and the target(s) of the mission.

An SVC (service) mission consists of the refueling of a strike aircraft by an orbiting refueling tanker. The properties of an SVC mission include the start time of the SVC, the length of time that the servicing takes (called the duration of the servicing), and the amount of fuel disbursed to the strike aircraft.

An RFL (refueling) mission consists of a tanker aircraft flying a certain route in order to service one or more strike aircraft.

An STN (station) mission consists of a tanker aircraft stationing itself, by orbiting for a certain period of time, at a certain location in order to service one or more strike aircraft.

An STK (strike) mission consists of a strike aircraft striking its target. Properties of the STK mission include the identifier of the target, target location, and the time of the target strike.

#### Granularity

If a system has a knowledge/database that contains a voluminous amount of information, the system must have a method of selecting the appropriate information to present in response to user requests if the response is not well constrained by the request itself. A factor that can play an important role in selecting appropriate information is that of granularity and, indeed

**Table 18.1** Granularity

| *Node Level* | *Very Coarse* | *Coarse* | *Default* | *Fine* | *Very Fine* |
|---|---|---|---|---|---|
| Parents | name | name<br>major | name<br>major | name<br>major | name<br>major<br>minor |
| Node | name | name<br>major | name<br>major<br>minor | name<br>major<br>major | name<br>minor<br>major |
| Children | | name | name | name<br>major | name<br>major<br>minor |

CUBRICON uses the concept of granularity for this purpose. The CUBRICON design includes two types of granularity:

1. Detail granularity: The first type is a measure of degree of detail and includes measures called "coarse," "medium," and "fine." The properties of each task type are divided into two groups: major and minor. These two groups roughy separate the properties into the more important properties of a mission and the less important properties, respectively. The name of a task is treated as special.

   Table 18.1 shows how the levels of detail granularity are defined in the current CUBRICON prototype. Each column of the table, other than the first, defines a granularity level or type. For each granularity type, the corresponding column shows the information to be presented about the parent tasks, the task itself, and the children tasks. For example, for coarse granularity and a task X, the name property and all major properties of the parent are presented, the name and major properties of X itself, and the names of the children tasks are presented.

2. Scope granularity: The second type is a scope concept and is measured on a scale of "local" to "global." For any task type, scope granularity is relative to the task node in the hierarchy. "Local" refers to information about a particular task node itself, and nonlocal granularity would include information about one or more child, parent, grandchild, grandparent, etc., nodes.

## Multimedia Output Composition

This section discusses the CUBRICON generation component that composes multimedia output for *collections of compound space/time-dependent activities*. For the air force mission planning domain of this project, the compound activities are OCA missions. These missions are composed of sub-missions, with their flight paths forming *activity sequences*.

The presentation of a collection of compound activities is subdivided into two parts by the multimedia output grammar: an *introduction* and a *main body*. The introduction provides summary highlights of the activities, and the main body provides a detailed presentation of the activities. The main body of the discourse consists of the presentation of the main space/time-dependent activities and their related activities. The introduction uses *coarse granularity*, the main body of the presentation uses *medium granularity*, and ancillary activities are presented using *fine local-only granularity*.

For each OCA mission presentation, the main activity sequence is the flight path traversal. A path traversal sequence is composed of activities, each of which consists of traversing a leg (line segment) of the (polygonal) path. The arrival at the endpoint of the leg is an event. This terminology is similar to that of Activity-on-Edge Networks. We call the endpoints of each of the line segments "waypoints," the common military terminology. Each of these segment traversals takes place in both space and time: The waypoints of each segment have coordinates in terms of latitude/longitude, and each waypoint has an associated time at which the aircraft traversing the path is scheduled to reach the waypoint. The segments of each path are sequenced by their common endpoints; that is, the last (second) endpoint of line segment $S$ is the first endpoint of line segment $S + 1$.

Important ancillary activities occur during the main space/time-dependent activity sequence. The ancillary activities are submissions or subtasks of the main OCA mission, including tasks such as the striking of targets and the airborne refueling of strike aircraft. After each leg (segment) of the path is presented, the multimedia generator determines whether there are any ancillary activities whose occurrence time or start time precedes the occurrence time of the next waypoint. If so, these ancillary activities are presented in order of their time of occurrence as represented in the knowledge base. Otherwise, the next leg or segment of the path is presented. Figure 18.6 shows the color graphics display after two flight paths have been presented as part of a multimedia presentation.

After each leg-transition activity is presented, the relevant and timely ancillary activities are presented. As indicated previously, for an OCA mission presentation these ancillary activities are the subtasks of striking the targets and the refueling of the strike aircraft. For each ancillary activity, local granularity is used to select the information to present. The information is generated in multimedia language.

The CUBRICON natural language generator composes and presents both spoken and written (on the CRT) natural language. As discussed in the previous section, when graphic gestures and expressions are used, spoken natural language is generated so that the natural language and graphics are temporally synchronized. The system also produces a written version of its natural language utterances. During the multimodal presentation of a collection of complex activities such as mission plans, the written version of the natural language is presented in special dynamic text windows (one per OCA mission) on the color graphics CRT.

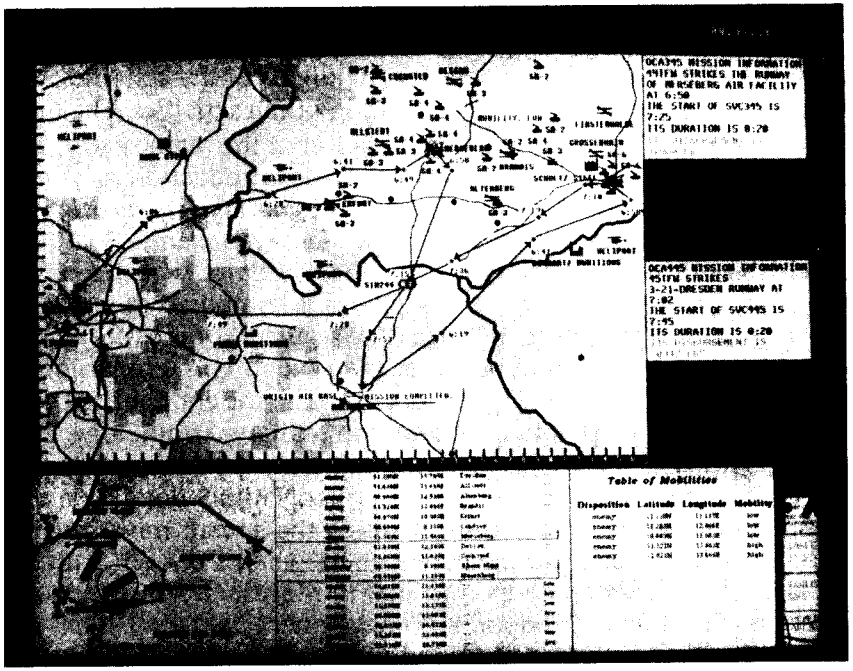

**Figure 18.6** Color graphics display after presentation of missions

Figure 18.6 shows the color graphics display after the conclusion of a multimodal presentation of a collection of space/time-dependent activities, which are OCA missions in this application.

## Spoken versus Visually Displayed Natural Language

An important feature of the CUBRICON system is that it distinguishes between the manner in which it generates spoken and visually displayed natural language. CUBRICON embeds deictic expressions with pointing gestures (e.g., "this SAM" with an accompanying pointing gesture to call the user's attention to the SAM system on one of the displays) in spoken output but not with NL printed to the screen, since a pointing or graphic gesture needs to be temporally synchronized with the corresponding verbal phrase, allowing for multiple deictic expressions within any individual sentence. The coordination between a pointing gesture and its co-referring verbal phrase is lost if it is embedded in printed text instead of speech, since it is difficult for a user to look in two places at the same time on the screen, that is, both at the natural language being printed to the screen and at a pointing gesture. When CUBRICON generates natural language to be visually displayed, definite descriptions are generated for noun phrases (with hopefully sufficient specificity to avoid ambiguous references) instead of simple demonstrative pronouns (with optional head nouns) and pointing gestures.

There are a number of different types of pointing gestures that CUBRICON uses, depending on the type of object being pointed to, the dialogue context, and the modality in which the object is visually presented. The primitives that CUBRICON combines for deictic expressions are speech, blinking, highlighting, graphic devices such as circling or boxing an item or region, "pointing text boxes," and flashing the border of a window (for pointing to a window).

## Example Multimedia Presentation Summary

The multimodal presentation of a collection of OCA missions is fairly lengthy and complex. The following summarizes such a presentation from the perspective of what a viewer sees and hears. Figure 18.6 shows the color graphics display after the conclusion of this multimedia presentation.

1. As an introduction to the presentation, for each OCA mission, its ID number, its package number (a package is a set of related missions), the origin (departure) airbase, and the OCA's submissions (strike and refueling missions) are summarized in speech and written language (on the Natural Language Interaction Window), accompanied by temporally synchronized pointing gestures to the corresponding items (as they are presented) on the mission form, which is on the monochrome display. For each OCA mission, a mission information window is initialized on the color graphics display, next to the relevant map window. It is used during the rest of the presentation to summarize important information in a written form.
2. The startup of each mission flight is signaled by the presentation of information about the aircraft departing the origin airbase in spoken natural language, with accompanying deictic and graphic gestures shown on the map display. Other relevant information includes the location of the origin airbase, the time of departure, and the unit that is flying the mission.
3. One by one, the segments making up the different (polygonal) flight paths are displayed in an animated manner on the map window so as to simulate simultaneous flight path traversal. Each flight path segment is presented at its appropriate time according to the time on waypoint for its second endpoint. For each flight path, an aircraft icon moves from waypoint to waypoint as a directed line segment grows to represent the particular leg of the flight path. As the aircraft reaches each waypoint, the time on waypoint is printed next to it.
4. For each mission flight, when the target of the mission is reached, it is identified via spoken natural language with synchronized deictic gestures consisting of blinking/highlighting its icon on the map window, highlighting the information on the form window, and any tables that include the information. The time on target is also presented. The target information is also summarized in the mission information window on the color graphics CRT.

5. For each mission, when the refueling location is reached on the map window, information about the refueling mission is presented in spoken natural language with synchronized simultaneous deictic gestures pointing to the relevant information representations on the various windows (i.e., map, form, and tables). The information is also summarized in the mission information window.
6. For each mission, when the aircraft arrives back at the origin airbase, the completion of the presentation is announced using speech output.

## 18.4 RELATED RESEARCH

In this section, we discuss related research in which artificial intelligence is applied in the following areas: intelligent multimedia interfaces, multimedia authoring, multimedia information and document storage and retrieval, and multimedia conferencing and decision making.

### 18.4.1 Intelligent Multimedia Interfaces

Multimedia dialogue for human-computer interaction has been a focus for several research groups. We consider both input and output in the following paragraphs.

For the understanding of multimedia input, our CUBRICON project focused on the understanding of natural langauge accompanied by simultaneous coordinated pointing gestures, particularly the problem of referent identification. Related work includes the development of the TEMPLAR system [41] at TRW, XTRA [42,43] at the University of Saarbrucken, and the IMAGE system [44] at the ATR Communication Systems Research Lab. The TEMPLAR system seems to provide only for a pointing gesture to substitute for a natural language definite reference within a natural language sentence, rather than also allowing a pointing gesture to be used in combination with an NL reference. The CUBRICON approach is closer to that of Kobsa and colleagues with the XTRA system, in accepting dual-modality input and applying several knowledge sources for referent identification. The IMAGE system is an illustrated map guidance system with an interface that enables the user to use voice and hand pointing gestures for spatial layout and/or positioning during input. The IMAGE system uses a DataGlove for hand pointing gestures, instead of the mouse device used for pointing gestures in the other systems mentioned above.

The automated composition and generation of multimedia dialogue output and presentations from knowledge base representations has been addressed by several efforts: CUBRICON [10,11,12,13], Integrated Interfaces [45,46], COMET (Coordinated Multimedia Explanation Testbed) [47,48], WIP [49,50], SAGE [51], and Maybury [52]. The CUBRICON effort was discussed in Section 18.3. The Integrated Interfaces system is based on models of the objects of the application and the interface and their classes, relationships, behaviors, and behavior effects. The prototype is capable of producing presentations that include map graphics, natural language, menus, and forms for a navy situation database-reporting application.

The COMET, WIP, SAGE, and Maybury projects are addressing the issue of composing and generating coordinated graphics and text in such a way that they complement one another. Planning paradigms are responsible for deciding what to say and how to say it. The presentation planners can be considered to construct and organize presentation material to achieve communicative goals. COMET is designed to generate explanations for an expert system concerning how to carry out field maintenance and repair procedures for military radio equipment. Its architecture is based on separate communicating media generators for text and graphics. The SAGE system is designed to produce explanations of changes in quantitative models used in programs such as spreadsheets, database, and scheduling tools. Currently, the WIP system is being designed to generate instructions for a user to properly manipulate real-world devices, such as an espresso machine. In the WIP system, generation is controlled by a set of generation parameters such as target audience, presentation objective, resource limitations, and target language.

### 18.4.2 Multimedia Information/Document Storage and Retrieval

Information/document storage and retrieval issues are fundamental to multimedia document navigation and multimedia document authoring. First, since the components of a hypertext or multimedia document are linked, the lines between documents become blurred; that is, the information packets that comprise the chapters or sections of a multimedia document can be considered as separate documents or as components of the same (or several) document(s). Following a hypertext or multimedia document link while viewing or navigating a multimedia document is a simple form of associative information retrieval.

Second, a critical component of the multimedia or hypertext authoring process is the determination of what information to include in a multimedia document and what links to establish. Again, a key function that needs to be provided to support the multimedia author is the location and retrieval of relevant information, both from documents that the author(s) has (have) created or from the works of other authors.

Therefore, we review related research in the applications of artificial intelligence technology to multimedia information storage and retrieval in this subsection from the perspective of viewers (or readers or navigators) and authors (creators) of multimedia documents.

Limited forms of artificial intelligence technology have recently been incorporated into hypertext and multimedia document manipulation tools. For example, KnowledgePro from Knowledge Garden, Inc., is a

xpert system development tool for the IBM PC that includes a hypertext acility. The links provided by KnowledgePro can trigger rules that query he user and guide document navigation.

Researchers at the NASA Ames Research Center are addressing the prob- em of document management and maintenance in the design and devel- opment of their Computer Integrated Documentation (CID) system 2,3,53]. Their approach is to design and build an intelligent problem- driven context-sensitive tool that interacts with and learns from users, and ises interaction media including intelligent hypertext, multimedia, and irtual environments. The focus of this work is on combining conventional ndexing, hypertext, and knowledge-based systems to develop semantic ontext-sensitive indexing and information retrieval. The objective is to mprove the precision and recall for document retrieval. Their approach lepends on the use of contexts which define mappings from descriptors to locument referents. The NASA researchers call their documents with asso- iated indexing knowledge "active documents," since they can be consid- red as knowledge-based systems in the sense that they can present relevant nformation in an appropriate format based on context knowledge.

Cooperative research by the NASA Ames Research Center and the Center or Design Research at Stanford University is focusing on the development f methods for capturing and storing design knowledge and a knowledge- ased interface for information retrieval [54,55]. The goal is to facilitate the euse of design knowledge and information.

Recent research in the application of artificial intelligence technology to locument/data indexing and search methods include new approaches to ext analysis for automating the key phrase indexing process [56], deductive ypermedia technologies [57], and expert system approaches to image lassification and retrieval [58,59,60].

Research at the MIT Media Laboratory is addressing the problem of epresenting the content of multimedia information, with emphasis on the roblem of data entry or video logging [61]. The representation of the nformation must be able to support all aspects of the process of video locument authoring: logging the footage into the archive, displaying nformation about the footage, retrieving it, and inserting it into the new ideo document. An iconic user interface, called the Director's Workshop, s being developed to support the process of describing multimedia infor- nation for later retrieval and resequencing by automatic presentation ystems.

### 8.4.3 Multimedia Conferencing and Decision Making

Computer-supported cooperative work (CSCW) [62,63,64,65,66] has at- racted dramatically increasing attention during the past several years and s an area in which multimedia systems are being applied to support group onferencing and decision making [5,6,67,68]. Just as we have seen the pplication of artificial intelligence to the problem of providing single-user decision support systems, artificial intelligence is now being applied to the more difficult problem of supporting collaborative work by groups of decision makers. Systems supporting collaborative or group decision mak- ing are primarily distributed systems, and, therefore, much of the artificial intelligence research in this area is classified as "distributed artificial intel- ligence" [69,70].

Moulin, Chaib-draa, and Cloutier [71] are working on a multi-agent system in which several artificial and human agents are able to interact together and also work individually and/or jointly on a planned course of action. Their work is based on a model of speech acts and ordinary acts using communication/action structures. Decision spaces are used to model agents' states and their relations: effective states, potential states, inten- tions, and commitments. Each artificial agent includes three layers: an operational component which builds plans; a tactical component which reasons about the agent's motivations, intentions, and goals; and a strategic component which reasons about the agent's commitments in relation to other agents' beliefs and intentions.

Kaye and Karam [72] are designing and implementing a system to support office workers using an approach based on embedding office knowledge in a network of distributed cooperating knowledge-based "assistants" and servers. This distributed system incorporates both factual and procedural knowledge and is capable of making use of existing conventional office technology.

## 18.5 FUTURE DIRECTIONS

The development of hardware user interface devices is advancing rapidly, providing ever-more sophisticated 3-D graphics and animation systems, video manipulation and presentation systems, 3-D sound and speech production systems, speech recognition systems, DataGlove devices, Datasuits, eye-tracking devices, stereo goggles, machine vision systems to capture and interpret user body language and facial expressions, and other devices that will contribute to creating virtual realities for human-computer interaction.

The problems for the future lie in the integration and intelligent use and control of the various devices and media: the problems will be with software development, not hardware.

The critical areas of artificial intelligence technology in which future significant progress is needed include:

**Self-adaptive systems:** As systems become more complex with regard to their ability to use combinations of larger suites of communication modes and media, adapting, tailoring, and porting any given system to another user, user group, user environment, or application will become increasingly difficult. Therefore, it becomes increasingly important for systems to automatically adapt or modify themselves appropriately, while

giving the user a feeling of control over the system and interaction process and providing a stable interaction environment.

**Multimedia input understanding:** Significant progress needs to be made in the coordination and interpretation of complex multimedia input, composed of media such as spoken natural language, facial and other body language expressions, eye-tracking information, and pointing gestures. Even combining two communication modes poses difficult problems, and the history of research into combining multiple input modes is relatively brief. Furthermore, computer understanding of some of the individual communication modes is still not satisfactorily accomplished. For example, after many years of research, natural language understanding is still a problem to which there is no general satisfactory solution, and it remains one of the most challenging fields within artificial intelligence.

**Intelligent automated multimedia output generation:** As the volume and types of information and presentation types become more numerous and varied, it will become increasingly important for computer systems to make automated knowledge-based decisions regarding information presentation to users. If we consider the status of the technology for composing and temporally and spatially coordinating information presentations using just two media, natural language and graphics, we find that the problem is difficult and the technology is just in its infancy. As more media and modalities are added to the presentation suite, the problems become more severe and the need for research becomes more significant.

Such intelligent automated multimedia information presentation has application in many areas, including multimedia or hypermedia document presentation, information retrieval from multimedia databases and knowledge bases, and explanation subsystems of user help facilities.

Some of the subproblems of multimedia output generation include:

- ☐ Selection of media and apportionment of information content among the media used for information presentation.
- ☐ Coordination of the media with respect to both space and time.
- ☐ Consistency of selection, composition, and generation across presentations.

**Knowledge-based development tools for multimedia systems:** As more media become viable and available for computer systems, the need for more sophisticated development tools becomes more critical. For example, high-quality authoring systems for hypermedia or multimedia presentation/document authoring are needed as well as development tools for multimedia human-computer interfaces. Generic intelligent automated multimedia composition and generation technology, mentioned above, could play an important role in alleviating some of the development problems.

**Hypermedia document systems:** As mentioned above, the area of hypermedia/multimedia document authoring and presentation systems requires better system development tools and intelligent automated multimedia presentation technology. Another area in which hypermedia document technology could benefit is in the area of "smart links," that is, hypertext/media links which have some decision-making ability to lead the viewer/reader to appropriate information based on factors such as what the system knows the viewer has already "read," the viewer's goals and objectives for viewing the material, and the viewer's background and level of expertise in relevant fields.

**Intelligent interfaces:** Systems are becoming capable of manipulating ever-more sophisticated object types because of increased variety in the storage media used in databases and knowledge bases. Research and development is already in progress in the area of intelligent interfaces, but the need for more highly intelligent interfaces will increase to relieve the user of the need to know the increasing capabilities of information manipulation computer systems.

**Computer supported cooperative work:** Collaboration is pervasive and complex, and it brings an associated suite of difficult problems to each of the collaborators in areas such as the decomposition of tasks into subtask assignments, scheduling, coordination, communication, planning, searching and retrieving information, sharing resources, integrating individual collaborators' products into group products, and understanding one another's goals, plans, activities, and accomplishments. This field will benefit significantly from the application of artificial intelligence research and development to CSCW systems. Artificial intelligent agents, for example, could provide assistance and guidance to their human counterparts and relieve them of their more mundane and time-consuming tasks.

## 18.6 SUMMARY

This chapter has focused on applications of artificial intelligence to multimedia systems. We presented a concept and design for a multimedia integrated workstation environment with a human-computer interface designed as a supportive intelligent agent with the ability to communicate and make presentations in coordinated multiple media/modalites.

The design and functionality for an implemented prototype system called CUBRICON, was discussed. CUBRICON has several unique features:

- ☐ CUBRICON accepts and understands natural language input accompanied by simultaneous pointing gestures. CUBRICON allows a variety of object types to be targets of point gestures and accepts a variable number of multimodal phrases within any sentence. CUBRICON can also use natural language inputs to disambiguate corresponding point gestures, and vice versa. CUBRICON also handles certain types of ill-formed multimodal inputs.
- ☐ CUBRICON composes and generates relevant output to the user in coordinated multiple modalities. CUBRICON selects appropriate in

formation to output to the user based on the user request, relevance to the user's task, relevance to the dialogue, and consistency of output displays. CUBRICON selects appropriate output media/modalities based on the characteristics of the information to be expressed as well as task and dialogue context. The output modalities are used in a highly integrated manner. Multimedia outputs, especially speech and accompanying graphics, are temporally synchronized. The system distinguishes between spoken and written natural language output and composes such natural language outputs appropriately.

- ☐ CUBRICON provides intelligent automatic management of windows in a dual-monitor environment. This includes a method for determining window importance that is used to decide which windows to remove when display space is needed for other windows.
- ☐ CUBRICON includes several knowledge sources to support its decision-making processes. These knowledge sources include a dialogue model, user-task model, and a knowledge base of interface and task-specific information.

Related research is being conducted at various institutions into the application of artificial intelligence to multimedia interfaces, multimedia authoring, multimedia information and document storage and retrieval, and multimedia conferencing and decision making. We reviewed some of the research in these areas.

The application of artificial intelligence to multimedia systems shows promise of significant future benefit in producing systems that provide a natural multimedia language for human-computer interaction, provide more assistance and support for users, relieve the user of the burden of managing the interface, and are adaptable to tasks, users, and context.

## 18.7 REFERENCES


1. Clark, D. R. The Demise of Multimedia. *IEEE Computer Graphics and Applications*. vol. 11, no. 4. 1991. pp. 75–80.
2. Boy, G. A. *Computer Integrated Documentation*. NASA Technical Memorandum 103870. 1991.
3. Boy, G. A. Semantic Correlation in Context: Application in Document Comparison and Group Knowledge Design. *Proceedings of the AAAI Spring Symposium on Cognitive Aspects of Knowledge Acquisition*. 1992.
4. Cornell, M., Suthers, D., and Woolf, B. Using "Live Information" in a Multimedia Framework. *Proceedings of the AAAI-91 Intelligent Multimedia Interfaces Workshop*. 1991. pp. 93–98.
5. Bly, S. A., Harrison, S. R., and Irwin, S. Media Spaces: Video, Audio, and Computing Environment. *Communications of the ACM*. vol. 36, no. 1. 1993. pp. 28–47.
6. Fish, R. S., Kraut, R. E., Root, R. W., and Rice, R. E. Video as a Technology for Informal Communication. *Communications of the ACM*. vol. 36, no. 1. 1993. pp. 48–61.
7. Grimes, J. and Potel, M. Guest Editors' Introduction: Multimedia—It's Actually Useful! *IEEE Computer Graphics and Applications*. vol. 11, no. 4. 1991. pp. 24–25.
8. Grice, H. P. Logic and Conversation. In P. Cole and J. L. Morgan (eds.), *Syntax and Semantics, vol. 3: Speech Acts*. Academic Press. 1975. pp. 41–48.
9. Neal, J. G., Bettinger, K. E., Byoun, J. S., Dobes, Z., and Thielman, C. Y. An Intelligent Multimedia Human-Computer Dialogue System. *Proceedings of the Workshop on Space Operations, Automation, and Robotics (SOAR-88)*. 1988. pp. 245–251.
10. Neal, J. G., Dobes, Z., Bettinger, K. E., and Byoun, J. S. Multimodal References in Human-Computer Dialogue, *Proc. AAAI-88*. 1988. pp. 819–823.
11. Neal, J. G., Thielman, C. Y., Dobes, Z., Haller, S. M., and Shapiro, S. C. Natural Language with Integrated Deictic and Graphic Gestures. *Proc. of the DARPA Speech and Natural Language Workshop*. 1989. pp. 410–423.
12. Neal, J. G., Shapiro, S. C., Thielman, C. Y., Lammens, J. M., Funke, D. J., Byoun, J. S. Dobes, Z., Glanowski, S., Summers, M., Gucwa, J. R., and Paul, R. *Final Report for the Intelligent Multi-Modal Interfaces Project*. RADC Technical Report TR-90-128. 1990.
13. Neal, J. G., and Shapiro, S. C. Intelligent Multi-Media Interface Technology. In J. W Sullivan S. W. Tyler (eds.), *Intelligent User Interfaces*. Addison-Wesley. 1991. pp. 11–44.
14. Shapiro, S. C. The SNePS Semantic Network Processing System. In Findler, (ed.), *Associative Networks—The Representation and Use of Knowledge by Computers*. Academic Press. 1979. pp 179–203.
15. Shapiro, S. C. *SNePS User's Manual*. The SNePS Implementation Group. Computer Science Deptartment, SUNY at Buffalo, NY. 1981.
16. Shapiro, S. C., and Rapaport, W. SNePS Considered as a Fully Intensional Propositiona Semantic Network. *Proc. AAAI-86*. In G. McCalla and N. Cercone (eds.), *Knowledg Representation*. Springer-Verlag. 1987. pp. 278–283.
17. Shapiro, S. C. Generalized Augmented Transition Network Grammars for Generation fron Semantic Networks. *AJCL*. vol. 8, no. 1. 1982. pp. 12–25.
18. Maida, A. S., and Shapiro, S. C. Intensional Concepts in Propositional Semantic Networks In R. J. Brachman H. J. Levesque (eds.), *Readings in Knowledge Representation*. Morgai Kaufmann. 1985. pp. 169–190.
19. Shapiro, S. C., and Neal, J. G. A Knowledge Engineering Approach to Natural Languag Understanding. *Proc. ACL*. 1982. pp. 136–144.
20. Neal, J. G., and Shapiro, S. C. Knowledge Representation for Reasoning About Language In J. C. Boudreauz, B. W. Hamill, and R. Jernigan (eds.), *The Role of Language in Problem Solving*. Springer-Verlag. 1986. pp. 27–47.
21. Neal, J. G., and Shapiro, S. C. Knowledge Based Parsing. In L. Bolc (ed.), *Natural Language Parsing Systems*. Springer-Verlag. 1987. pp. 49–92.
22. Shapiro, S. C., Martens, J., and McKay, D. Bidirectional Inference. *Proc. of the Cognitiv Science Society*. 1982. pp. 90–93.
23. McKay, D. P., and Shapiro, S. C. MULTI—A LISP-Based Multiprocessing System. *Conferen Record of the 1980 LISP Conference*. Stanford University. 1980. pp. 29–37.
24. Shapiro, S. C. Using Nonstandard Connectives Quantifiers for Representing Deductio Rules in a Semantic Network. Paper presented at Current Aspects of AI Research, a semin held at the Electrotechnical Laboratory, Tokyo. 1979.
25. Miller, G. A. The Magical Number Seven Plus or Minus Two. *Psychological Review* 63. 195 pp. 81–97.
26. Andriole, S. J. Graphic Equivalence, Graphic Explanations, and Embedded Process Mo eling for Enhanced Process Modeling for Enhanced User-System Interaction. *IEEE Tran on Systems, Man, and Cybernetics*. vol. 16, no. 6. 1986
27. Grosz, B. J. Discourse Analysis. In D. Walker (ed.), *Understanding Spoken Language*. Elsevi 1978. pp. 229–345.
28. Grosz, B. J. The Representation and Use of Focus in a System for Understanding Dialog In B. J. Grosz, K. S. Jones, and B. L. Webber (eds.), *Readings in Natural Language Processin* Morgan Kaufmann. 1986. pp. 353–362.

29. Sidner, C. L. Focusing in the Comprehension of Definite Anaphora. In B. J. Grosz, K. S. Jones, B. L. Webber (eds.), *Readings in Natural Language Processing*. Morgan Kaufmann. 1986. pp. 353–362.
30. Grosz, B. J., and Sidner, C. L. Discourse Structure and the Proper Treatment of Interruptions. *Proc. of IJCAI*. 1985. pp. 832–839.
31. Grosz, B. J. Focusing and Description in Natural Language Dialogues. In A. Joshi, B. Webber, and I. Sag (eds.), *Elements of Discourse Understanding*. Cambridge University Press. 1981. pp.84–105.
32. Carberry, S. First International Workshop on User Modeling. *AI Magazine*. vol.8, no.3. 1987. pp.71–74.
33. Kobsa, A., and Wahlster, W. (eds.). *Computational Linguistics. Special Issue on User Modeling*. MIT Press. 1988.
34. Kobsa, A., and Wahlster, W. (eds.). *User Models in Dialog Systems*. Springer-Verlag. 1989.
35. Kass, R., and Finin, T. General User Modeling: A Facility to Support Intelligent Interaction. In J. W. Sullivan and S. W. Tyler (eds.), *Intelligent User Interfaces*. ACM Press/Addison-Wesley. 1991. pp. 111–128.
36. Wahlster, W. User and Discourse Models for Multimodal Communication. In J. W. Sullivan and S. W. Tyler (eds.), *Architectures for Intelligent Interfaces: Elements and Prototypes*. Addison-Wesley/ACM Press. 1991.
37. Malone, T. W. How Do People Organize Their Desks? Implications for the Design of Office Automation Systems. *ACM Transactions on Office Information Systems* 1(1). 1983. pp. 99–112.
38. Funke, D. J., Neal, J. G., and Paul, R. D. An Approach to Intelligent Automated Window Management, *International Journal of Man-Machine Studies*. 1993.
39. Bly, S. A., and Rosenberg, J. K. A Comparison of Tiled and Overlapped Windows. *CHI '86 Proceedings*. 1986. pp. 101–106.
40. Billingsley, P. A. Taking Panes: Issues in the Design of Windowing Systems. In M. Helander (ed.), *Handbook of Human-Computer Interaction*. Elsevier. 1988. pp. 413–436.
41. Press, B. The U.S. Air Force TEMPLAR Project Status and Outlook. *Western Conf. on Knowledge-Based Engineering and Expert Systems*. 1986. pp. 42–48.
42. Allgayer, J., Jansern-Winkeln, R., Reddig, C., and Reithinger, N. Bidirectional Use of Knowledge in the Multimodal NL Access System XTRA. *Proc. of IJCAI-89*. 1989. pp. 1492–1497.
43. Kobsa, A., Allgayer, J., Reddig, C., Reithinger, N., Schmauks, D., Harbusch, K., and Wahlster, W. Combining Deictic Gestures and Natural Language for Referent Identification. *Proc. of the 11th International Conference on Computational Linguistics*. 1986.
44. Takahashi, T., Hakata, A., Shima, N., and Kobayashi, Y. Unifying Voice and Hand Indication of Spatial Layout. *SPIE vol. 1198. Sensor Fusion II: Human and Machine Strategies*. 1989. pp. 346–353.
45. Arens, Y., Miller, L., and Sondheimer, N. Presentation Design Using an Integrated Knowledge Base. In J. W. Sullivan and S. W. Tyler (eds.), *Intelligent User Interfaces*. ACM Press/Addison-Wesley. 1991. pp. 241–258.
46. Arens, Y., and Hovy, E. H. How to Describe What? Towards a Theory of Modality Utilization. *Proceedings of the 12th Cognitive Science Conference*. 1990.
47. Feiner, S. K., and McKeown, K. R. Automating the Generation of Coordinated Multimedia Explanations. *IEEE Computer*. vol. 24, no. 10. 1991. pp. 33–41.
48. Feiner, S. K., and McKeown, K. R. Coordinating Text and Graphics in Explanation Generation. *Proceedings of the 8th National Conference on Artificial Intelligence*. 1990. pp. 442–449.
49. Wahlster, W., Andre, E., Graf, W., and Rist, T. Designing Illustrated Texts: How Language Production is Influenced by Graphics Generation. *Proceedings of the 5th Conference of the EACL*. 1991. pp. 8–14.
50. Wahlster, W., Andre, E., Bandyopadhyay, S., Graf, W., and Rist, T. WIP: The Coordinated Generation of Multimodal Presentations from a Common Representation. In A. Ortony, J. Slack, and O. Stock (eds.), *A.I. and Cognitive Science Perspectives on Communication*. Springer-Verlag. 1991.
51. Roth, S. F., Mattis, J., and Mesnard, X. Graphics and Natural Language as Components of Automatic Explanation. In J. W. Sullivan and S. W. Tyler (eds.), *Intelligent User Interfaces*. ACM Press/Addison-Wesley. 1991. pp. 207–240.
52. Maybury, M. T. Planning Multimedia Explanations Using Communicative Acts. *Proceedings of the 10th National Conference on Artificial Intelligence*. 1991.
53. Mathe, N., and Boy, G. The Computer Integrated Documentation Project: A Merge of Hypermedia and AI Techniques. *Proceedings of SOAR '92*. 1992.
54. Baudin, C., Gevins, J., Mabogunje, A., and Baya, V. A Knowledge-Based Interface for Design Information Retrieval. *Proceedings of the AAAI-91 Intelligent Multimedia Interfaces Workshop*. 1991. pp. 133–140.
55. Baudin, C., Sivard, C., and Zweben, M. Recovering Rationale for Design Changes: A Knowledge-Based Approach. *Proceedings IEEE*. 1990.
56. Driscoll, J., Rajala, D., Shaffer, W., and Thomas, D. The Operation and Performance of an Artificially Intelligent Keywording System. *Information Processing and Management*. vol. 27, no. 1. 1991. pp. 43–54.
57. Parsaye, K., Chignell, M., and Khoshafian, S. *Intelligent Databases: Object-Oriented, Deductive Hypermedia Technologies*. John Wiley & Sons. 1989.
58. Ragusa, J. M., and Orwig, G. Expert Systems and Imaging: NASA's Start-Up Work in Intelligent Image Management. *Journal of Expert Systems*. vol. 3. 1990. pp. 25–30.
59. Ragusa, J. M., and Orwig, G. Attacking the Information Access Problem with Expert Systems. *Journal of Expert Systems*, vol. 4. 1990. pp. 26–32.
60. Ragusa, J. M. and Heard, A. Intelligent Multimedia Interfaces: Research Issues and Some Sample Applications. *Proceedings of the AAAI-91 Intelligent Multimedia Interfaces Workshop*. 1991. pp. 162–172.
61. Davis, M. E. Director's Workshop: Semantic Video Logging with Intelligent Icons. *Proceedings of the AAAI-91 Intelligent Multimedia Interfaces Workshop*. 1991. pp. 122–132.
62. Special Section on Computer-Supported Cooperative Work. *Communications of the ACM*. vol. 34, no. 12. 1991.
63. *CSCW '88: Proceedings of the Conference on Computer Supported Cooperative Work*. ACM Press. 1988.
64. *CSCW '90: Proceedings of the Conference on Computer Supported Cooperative Work*. ACM Press. 1990.
65. *CSCW '92: Proceedings of the Conference on Computer Supported Cooperative Work*. ACM Press. 1992.
66. Ellis, C. A., Gibbs, S. J., and Rein, G. L. Groupware: Some Issues and Experiences *Communications of the ACM*. vol. 34, no. 1. 1991. pp. 38–58.
67. Ishii, H., and Miyake, N. Toward an Open Shared Workspace: Computer and Video Fusior Approach of TeamWorkStation. *Communication of the ACM*. vol. 34, no. 12. 1991. pp 36–51.
68. Francik, E., Rudman, S. E., Cooper D., and Levine, S. Putting Innovation to Work: Adoptior Strategies for Multimedia Communication Systems. *Communications of the ACM*. vol. 34 no. 12. 1991. pp. 52–63.
69. Bond, A., and Gasser, L. *Readings in Distributed Artificial Intelligence*. Morgan Kaufman 1989.
70. Datta, A. Cooperative Problem Solving in Distributed Decision Making Contexts. *Proceed ings of the 1991 IEEE International Conference on Systems, Man, and Cybernetics*. 1991. pp 2085–2090.
71. Moulin, B., Chaib-draa, B., and Cloutier, L. A Multi-Agent System Supporting Cooperativ Work Done by Persons and Machines. *Proceedings of the 1991 IEEE International Conferenc on Systems, Man, and Cybernetics*. 1991. pp. 1889–1893.
72. Kaye, A. R., and Karam, G. M. Cooperating Knowledge-Based Assistants for the Office *ACM Transactions on Office Information Systems*. vol. 5, no. 4. 1987. pp. 297–326.

## 18.8 FOR FURTHER READING


1. *Communications of the ACM: Special Issue on HyperText.* vol. 31, no. 7. 1988.
2. Cheikes, B. A., and Webber, B. L. The Design of a Cooperative Respondent. In J. W. Sullivan S. W. Tyler (eds.), *Architectures for Intelligent Interfaces: Elements and Prototypes.* Addison-Wesley/ACM Press. 1990.
3. Conklin, J. Hypertext: An Introduction and Survey. *Computer.* vol. 20, no. 9. 1987. pp. 17–41.
4. Egan, D. E., Remde, J. R., Gomez, L. M., Landauer, T. K., Eberhardt, J., and Lochbaum, C. C. Formative Design-Evaluation of SuperBook. *ACM Transactions On Information Systems.* vol. 7, no. 1. 1989. pp. 30–58.
5. Hollan, J., Miller, J. R., Rich, E., and Wilner, W. Knowledge Bases and Tools for Building Integrated Multimedia Intelligent Interfaces. In J. W. Sullivan and S. W. Tyler (eds.), *Architectures for Intelligent Interfaces: Elements and Prototypes.* Addison-Wesley/ACM Press. 1990.
6. Kaplan, S. J. Cooperative Responses From a Portable Natural Language Database Query System. In M. Brady (ed.), *Computational Models of Discourse.* MIT Press. 1982.
7. Reithinger, N. Generating Referring Expressions and Pointing Gestures. In G. Kempen (ed.), *Natural Language Generation.* Nijhoff. 1987. pp. 71–81.
8. Shapiro, S. C. Generation as Parsing from a Network into a Linear String. *AJCL.* 1975. pp. 45–62. Microfiche 33.
9. Shepard, S. J. A New Approach to Hypertext: MINDS. *AI Expert.* vol. 4, no. 9. 1989. pp. 69–72.
10. Simmons, R., and Slocum, J. Generating English Discourse from Semantic Networks. *CACM* 15:10. 1972. pp. 891–905.
11 Sullivan, J. W., and Sherman, W. T. (eds.). *Architectures for Intelligent Interfaces: Elements and Prototypes.* Addison-Wesley/ACM Press. 1990.