Intelligent Multimedia Interfaces

Intelligent Multimedia Interfaces

Edited by

Mark T. Maybury

MENLO PARK, CALIFORNIA
CAMBRIDGE, MASSACHUSETTS
LONDON, ENGLAND

AAAI Press / The MIT Press

445 Burgess Drive
Menlo Park, California 94025
USA

Copublished and distributed by The MIT Press,
Massachusetts Institute of Technology
Cambridge, Massachusetts and London England

ISBN 0-262-63150-4

Manufactured in the United States of America

Preface

This collection is an outgrowth of the American Association for Artificial Intelligence (AAAI) Workshop on Intelligent Multimedia Interfaces which took place at Anaheim, California in August of 1991. Multimedia interfaces are computer interfaces that communicate with users using multiple media (e.g., language, graphics, animations, video, non-speech audio) in multiple modalities (e.g., written text versus spoken language). *Intelligent* multimedia interfaces go beyond traditional hypermedia or hypertext environments to both process input and generate output in an intelligent or knowledge-based manner.

The purpose of the AAAI workshop was threefold: (1) to bring together researchers and practitioners to report on current advances in intelligent multimedia interface systems and their underlying theories, (2) to encourage scientific interchange among these individuals, and (3) to evaluate current efforts and make recommendations for future investigations. The workshop addressed a broad range of issues spanning the disciplines of artificial intelligence, computational linguistics, computer graphics, cognitive science, education/intelligent tutoring, software design, and information retrieval.

In addition to the many previous workshops on individual media (e.g., text generation, graphics generation), related workshops and collections have focused on intelligent user interfaces in general [Sullivan and Tyler 1991; Gray et al. 1993], multimedia interface design [Blattner and Dannenberg 1992], and multimedia communication [Taylor and Bouwhuis 1989]. This collection focuses specifically on those intelligent interfaces that exploit multiple media and modes to facilitate human-computer communication. As a consequence, this collection will be of interest to researchers and practitioners in computer science, artificial intelligence, computer-human interaction, cognitive science, and graphics design.

The book is organized into three sections: Automated Presentation Design; Intelligent Multimedia Interfaces; and Architectural and Theoretical Issues. The chapters in the first section focus on methods for the automatic design of multimedia presentations. Multimedia design involves a number of complex issues addressed by these papers including temporal coordination of multiple media, the relationship of textual and graphical generation, automatic design of graphics, and modality selection (e.g., realizing language as text or speech.) The chapters in the second section report on several investigations into systems that integrate multimedia input and generate coordinated multimedia output. These prototypes point the way to possible future systems that will enhance human-computer interaction. A final section considers knowledge sources and processes required for processing multiple media. These include the need to represent and reason about models of tasks and information, media, the user, and the discourse context. While research in this entire area is still in its formative stages, the individual contributors and I hope that this initial collection will help foster the scientific interchange and motivate necessary research to solve many of the remaining fundamental problems.

Acknowledgments

I would like to thank the American Association for Artificial Intelligence and MIT Press, in particular Mike Hamilton and Bob Prior, as well as The MITRE Corporation for their support and cooperation. I thank the authors and reviewers who worked hard to interrelate their chapters. This book would not have been possible without the administrative support of Susan Hanlon who spent endless hours typing, faxing, mailing, and proofing the many versions of the book. Finally and most importantly, for her continuous support while I edited the book on weekends and nights, I dedicate this collection to my wife, Michelle.

Mark Maybury
Bedford, Massachusetts
March, 1993

Contents

Section 3: Architectural and Theoretical Issues / 277

Introduction

Mark T. Maybury

Abstract

Multimedia communication is ubiquitous in daily life. When humans converse with one another, we utilize a wide array of media to interact including spoken language, gestures, and drawings. We exploit multiple human sensory systems or modes of communication including vision, audition, and taction. Some media and modes of communication are more efficient or effective than others for certain tasks, users, or contexts (e.g., the use of speech to control devices in hand and eyes-busy contexts, the use of maps to convey terrain and cartographic information). Whereas humans have a natural facility for managing and exploiting multiple input and output media, computers do not. The ability of machines to interpret multimedia input and generate multimedia output would be a valuable facility for a number of key applications such as information retrieval and analysis, training, and decision support. This chapter introduces the need for intelligent multimedia interfaces, defines key terms and concepts, outlines the current state of the art, and describes the structure of this collection which addresses some remaining fundamental problems.

1. Need

Human abilities should be amplified, not impeded, by using computers, and the synergistic utilization of multiple media can support this amplification. If appropriate media are utilized for human computer interaction, there is the potential to (1) increase the bandwidth of information flow between human and machine (that is, the raw number of bits of information being communicated), and (2) improve the signal-to-noise ratio of this information (that is, the amount of useful bits conveyed). To achieve these potential gains, however, requires a better understanding of information characteristics, how they relate to characteristics of media, and how they relate to models of tasks, users, and environments. This goal is exacerbated by the proliferation of new interactive devices (datagloves and bodysuits, head mounted displays, three dimensional sound), the lack of standards, and a poor or at least ill-applied knowledge of human cognitive and physical capabilities with respect to multimedia devices. For example, some empirical studies [Krause, this volume] provide evidence that even well accepted applications of multimedia (e.g., the use of check marks and greying out in menus) can exacerbate rather than improve user performance. This motivates the need to understand the principles underlying multimedia communication. Understanding these principles will not only result in better models and interactive devices, but also lead to new tools for context-sensitive multimedia help, automated and semi-automated multimedia interface construction, and intelligent agents for multimedia information retrieval, processing, presentation, and authoring.

2. Definitions

We begin with a clarification of the terms multimedia and multimodal. By mode or modality we refer primarily to the human senses employed to process incoming information, e.g., vision, audition, taction, olfaction. We do not mean mode in the sense of purpose, e.g., word processing mode versus spread sheet mode. Additionally, we recognize medium, in its conventional definition, to refer both to the material object (e.g., paper, video) as well as the means by which information is conveyed (e.g., a sheet of paper with text on it). We would elaborate these definitions to include the possibility of layering so that a natural language mode might use written text or speech as media even though those media themselves rely on other modes.

Media and mode are related non-trivially. First, a single medium may support several modalities. For example, a piece of paper may support both language and graphics just as a visual display may support text, images, and video. Likewise, a single modality may be supported by many media. For example, the language modality can be supported visually (i.e., written language) and aurally (i.e., spoken language) – in fact spoken language can have a visual component (e.g., lip reading). Just as a single medium may support several modalities and a single modality may be supported by many media, many media may support many modalities, and likewise. For example, a multimedia document which includes text, graphics, speech, video, effects several modalities, e.g., visual and auditory perception of natural language, visual perception of images (still and moving), and auditory perception of sounds. Finally, this multimedia and multimodal interaction occurs over time. Therefore, it is necessary to account for the processing of discourse, context shifts, and changes in agent states over time.

Multimedia interfaces are computer interfaces that communicate with users using multiple media (e.g., language, graphics, animations, video, non-speech audio), sometimes using multiple modes, such as written text together with spoken language. "Intelligent" multimedia interfaces go beyond traditional hypertext or hypermedia environments in that they process input and generate output in an intelligent or knowledge-based manner. This area is multidisciplinary by nature, spanning the disciplines of at least artificial intelligence, computational linguistics, computer graphics, cognitive science, education/intelligent tutoring, software design, and information retrieval.

3. State of the Art

The state of the art in intelligent multimedia interfaces is exemplified by limited prototypes in narrow domains that are able to interpret a few kinds of input and generate limited forms of output. These systems often integrate or build upon single-media components that perform tasks such as spoken language recognition and generation or graphical design. For example, the state of the art includes the ability to interpret typed or spoken natural language utterances together with deictic mouse or dataglove gestures to resolve ambiguous references (e.g., "put that there") [Bolt 1980; Neal et al. 1989].

On the output side, the majority of work has investigated automated generation of single output media. In recent years a number of advances have been made in the area of linguistic realization (e.g., PENMAN [Mann 1983], MUMBLE [McDonald and Pustejovsky 1985], FUF [Elhadad et al. 1993]) and text planning (e.g., [McKeown 1985; Hovy 1988a; Moore 1989; Maybury 1990]). At the same time, others have made progress in graphical design. For example, mechanisms have been developed to design tables and charts [Mackinlay 1986b], network diagrams [Marks 1991ab], business graphics displays [Roth and Mattis 1991], and three dimensional explanatory graphics [Feiner 1985].

In addition, several laboratory prototypes have been developed that automatically generate coordinated multimedia presentations. For example, André et al. [this volume] describe WIP which presents and understands combinations of graphics, text, and pointing gestures (e.g., it can generate captioned visual instructions on how to operate an espresso machine). COMET (Columbia Operations and Maintenance Explanation Testbed) [Feiner and McKeown, this volume] automatically designs integrated textual and three dimensional graphical presentations to explain the operation and maintenance of an Army field radio. The integrated interfaces project is able to display Navy briefing information using maps, text, and tables [Arens et al. 1991] and SAGE automatically creates business graphics displays [Roth and Mattis 1991]. Finally, TEXPLAN [Maybury, this volume] provides narrated animations of directions over an object-oriented map using a collection of multimedia actions (e.g., speech acts, graphical acts) for media integration and control.

Other prototypes integrate multimedia input and output. For example, Burger and Marshall [this volume] describe an intelligent multimedia interface (AIMI) which can engage a user in a multimedia dialogue, for example, responding to a natural language query by automatically designing business-like graphics, which the user can then interact with or refer to. AIMI is able to choose alternative media to express information in an underlying KL-ONE-like knowledge base, for example, using non-speech audio to convey the speed, stage, or duration of an otherwise invisible process.

Despite the exciting possibilities suggested by these early prototypes, many fundamental questions remain only partially answered. This collection both reports on these early prototypes and begins to address some of these questions.

4. Key Remaining Problems: An Overview of the Book

From a system's standpoint, the key areas which require further investigation include the integration of multimedia input, the selection and coordination of multimedia output, and fuller knowledge of and better models for representing and reasoning about media and modes. From an architectual standpoint, we need to understand the infrastructure required to support and encourage progress in the field as well as fundamental questions such as: What are they key components?, What functionality do they need to support?, What is the proper flow of control?, and How should they interact? (e.g., serially, interleaved, co-constraining). Finally, we need to better understand from an empirical standpoint how well our integrated multimedia interfaces will function. This entails designing metrics and conducting evaluations. The book is organized around these issues into three sections: Section 1: Automated Presentation Design, Section 2: Intelligent Multimedia Interfaces, and Section 3: Architectural and Theoretical Issues.

4.1. Intelligent Multimedia Input and Output

The chapters in the first section of the book address the automatic generation of multimedia presentations, beginning with a survey by Roth and Hefley. Those in the second section of the book address interfaces which not only present but also interpret multimedia information.

At the most basic level, more investigation is needed in representing and reasoning about media. For input, current display-device technology remains cumbersome and low fidelity. Development and experimentation with new interactive devices (e.g., those replicating force feedback) and the integration of their multiple inputs (e.g., spoken language, gestures, and eye-trackers) is required. The integration of these must be done carefully to ensure synergistic coupling among multiple media. This is perhaps one of the least investigated areas, despite the widespread use of two input devices (i.e., keyboard and mouse). Koons et al. [this volume] describe an innovative approach to integrating speech, gaze, and gesture.

Multimedia output also requires further investigation. Many of the articles in the first and second sections of this book directly address this problem. Multimedia generation can be divided into the processes of content selection (i.e., choosing what to say), media allocation (choosing which media to say what in), media realization (choosing how to say items in a particular media), and media coordination. The design, realization, and coordination of text and speech, graphs, tables, pictures, maps, and forms offers a number of challenges. Key problems include the temporal coordination of multiple media, the relationship of textual and graphical generation, automatic design of graphics, and modality selection (e.g., realizing language as text or speech). The generation of multimedia presentations requires knowledge about the kind of information to display, the goal of the producer, the characteristics of the addressee, and the nature of the media (e.g., text versus graphics). Another issue concerns the degree of automation versus mixed initiative. Other issues concern whether or not systems save the history or structure of a presentation and if and how animations were connected to representations of abstract knowledge. The need for deep knowledge of designed graphics depends upon the intended use of the multimedia presentation (e.g., for teaching versus manual generation) and the environment in which it is used (e.g., interactive, static).

There are several common problems in allocating and coordinating media. These include the need for presentation balance, mutual-reference and the interaction between text and graphics, and the relationship between the characteristics of the information to be presented and the devices available for presentation. More sophisticated architectures may be required to control the design process (e.g., André et al.'s WIP [this volume] exploits two feedback loops, one after presentation design and one after realization, to help resolve inter- and intra-media synthesis problems). Related to the need to dynamically plan presentations is the choice between plan reuse, refinement, or replanning after a failed presentation. And when multiple choices among presentations are possible, one problem is the need for "goodness" metrics that, for example, measure the consistency and coherency of multimedia presentations.

A related issue concerns when, how, and why media are chosen to convey different types of information. Whereas some researchers take a practical approach to this problem, building systems that are based on reverse engineering of naturally-occuring presentations, others argue that media selection should be a machine-learned activity based on interaction with users, and still others argue that it requires empirical validation through observation of man-machine interactions. Related to this focus on empiricism is the need to provide statistical evidence that the additional machinery required to design and render more complex multimedia presentations is warranted by some pedagogic benefit, increase in efficiency, or increase in the effectiveness of accomplishing some task.

In summary, the capabilities of intelligent multimedia systems go beyond hypermedia to include the ability to interpret (possibly multimedia) questions and automatically design multimedia answers (e.g., WIP, COMET, AIMI), to deal with follow-up questions and make backward references (e.g., AIMI), and to post-edit presentations (e.g., COMET). Other areas which require further research include incorporating dialogue (e.g., context and turn-taking) into multimedia interfaces, more complex models of pedagogue, and a capability to provide diagnosis and advice giving as a user designs a presentation. A final possibility is tailoring multimedia presentations to individual user's psychological state, knowledge, abilities, attitudes and preferences, goals, and plans. The research results on reader adaptation in technical documentation, on user modeling in interactive computer systems, and on computer-aided tutoring systems, might also be relevant to this endeavor.

4.2. Architectural and Theoretical Issues

The last section of the book addresses issues concerning the architectures and empirical evaluation of intelligent multimedia interfaces. One of the primary concerns is what kinds of information and knowledge must be represented to support these systems, and how we should represent and reason about it. Necessary models include: (1) models of media (e.g., the characteristics, strengths and weaknesses), (2) models of the user (e.g., the ability to acquire, represent, and maintain useful data about user abilities (physical and cognitive), preferences, attention, and intentions from interactions with the interface), (3) modeling of dialogue history (e.g., the ability to automatically assimilate information from user interactions with the interface, and (4) modeling of the situation (e.g., the ability to automatically track system parameters such as load and available media, which can be used to influence interface decisions on input and output).

Questions remain as to how to acquire, represent, maintain, and exploit the models. Equally unspecified is the architectural relation of intelligent multimedia interface components – What is flow of control? Finally, there remains a need for facilities to integrate canned media with dynamically generated media.

Other issues are introduced in this collection by Roth and Hefley [this volume], Krause [this volume], and others, concerning metrics and methods for evaluating progress and capabilities in this area. First, it is necessary to more fully understand existing media. This includes representing media strengths and weaknessess in a standard manner, including the protocols that describe the kind of information these systems can use. This will demand standard terms, units of measurement, levels of performance, techniques of use, and so on, Equally important, however, is the need to match media to human (physical and cognitive) capabilities such as memory and attention. We will need to formulate metrics for time/quality tradeoff among media, and use these to judge among possible input and output facilities. Finally, we will require metrics for both glass box and black box evaluation of interface functionality to measure individual component effectiveness (e.g., timeliness and fidelity of generated media) as well as measuring overall interface effectiveness.

5. Conclusion

If successful, intelligent multimedia interfaces promise to enable systems and people to use media to their best advantage, in several ways. First, they can increase the raw bit rate of information flow between human and machine (for example, by using the most appropriate medium for information exchange). Second, they can facilitate human interpretation of information by helping to focus user attention on the most meaningful or relevant information. Third, they can use multiple media to more effectively allocate information across media during presentation. Finally, these investigations can provide explicit models of media to facilitate interface design so, for example, future interfaces can benefit from additional aspects of human communication that are currently ignored by current interfaces (e.g., speech inflections or hand gestures).

The goal of achieving context sensitivity will be limited only by the richness of models that can be created. In short, this area has the potential to improve the quality and effectiveness of interaction for everyone who communicates with a machine in the future. To achieve these benefits, however, we must overcome the remaining fundamental problems outlined above. The contributions in this book aspire to provide initial solutions.

6. Acknowledgments

I would like to thank all the workshop participants and authors for their ideas, many of which have been adopted above, especially Ed Hovy, Yigal Arens, Brad Goodman, John Burger, and Ralph Marshall. Finally, I thank Marc Vilain for the original inspiration for the workshop.

Section I

Automated Presentation Design

The papers in this first section raise in a concrete manner issues central to multimedia presentation design including: How can a system represent and reason about heterogeneous media in an integrated fashion? How should we select and apportion content to different media during design? How do we coordinate media? How can we ensure that given communicative goals are achieved by the resulting artifact? What is the relation between canned presentations and those that are dynamically designed and realized? authoring.

In the first chapter, Steven Roth and William Hefley place in historical and technical perspective the many investigations into intelligent multimedia presentation. They first consider the purposes of multimedia presentation systems and key functional requirements (e.g., content selection and presentation design, media apportionment and coordination). This leads to a consideration of their architectural structure and function, flow of control, and so on. They then consider the nature of the information contained in these systems. This is followed by a discussion of the range of functions of presentations and the implications this has on their underlying architectures, be they based on rules, constraints, rhetorical schema, plan operators, etc. They then consider various classes of presentation design knowledge associated with functions such as content selection, media and presentation technique selection, and presentation coordination. They conclude by providing a Human Computer Interaction (HCI) view of intelligent multimedia presentations which include concerns for evaluation metrics (e.g., usability as well as design complexity). A final section argues for the requirement for mechanisms to support interactive design, which entails defining vocabularies for specifying goals or tasks, methods for selecting among design alternatives, and possibilities for controlling rendering choices (e.g., color, fonts, orientation) or critiquing user designs.

The remaining chapters describe four systems which automatically design multimedia presentations, an extended TEXPLAN, WIP, COMET, and a visual repair prototype.

Chapter two describes an extension of a communicative act theory of multisentential text to multimedia presentations. In particular, building on the action-based view of communication advocated by Austin and Searle, the chapter defines linguistic, graphical, and physical actions as all being different methods of performing various communicative acts. These media-specific actions are abstracted into media-independent actions, called rhetorical actions, such as describe, compare, or explain. The chapter illustrates how a computational implementation of these ideas is able to represent and reason about multimedia actions in an integrated framework. This is exemplified by multimedia plans which were used in a object-oriented cartographic system for coordinated multimedia location identification and route exposition.

Immediately following this paper are two papers that detail the WIP system, which plans coordinated linguistic (English or German) and three-dimensional graphical displays (e.g., illustrated instructions for the operation of an espresso machine). The first chapter by Elisabeth André, Wolfgang Finkler, Winfried Graf, Thomas Rist, Anne Schauder, and Wolfgang Wahlster provides an architectural overview of WIP. Similar to TEXPLAN, WIP approaches multimedia design from a plan-based paradigm. Whereas TEXPLAN designs narrated animations of routes, WIP generates three dimensional graphics and embodies a constraint-based layout manager that is able to reject designs that do not fulfill layout constraints. Architecturally, WIP consists of two parallel processing cascades that enable the incremental design and realization of text (using tree adjoining grammars) and graphics. Further, individual media design and realization are interdependent processes that allow (graphical or textual) realization constraints to guide (graphical or textual) design goals, which in turn can constrain overall presentation or layout goals. This can happen both within and across media, exemplified by the generation of a cross-modal referring expression (e.g., "The on/off switch is located *in the upper left corner of the picture*").

The second WIP paper by Elisabeth André and Thomas Rist argues that traditional hierarchical planners are inadequate to handle complex interdependence of content determination, mode selection, and realization. They suggest that what is required are interleaved components which allow for revision and communication among one another. They outline mode preferences for information types such as: 1) prefer graphics for concrete information such as visual properties of objects or events involving physical objects, 2) prefer graphics for spatial information such as the location, orientation, composition, or movement of objects unless the emphasis is on minimizing errors in which case text is preferred, and 3) prefer text for quantification, negation, conditional, and causal relations if there is potential ambiguity. They also detail how mode decisions de-