The Knowledge Engineering Review, Vol. 11:1, 1996, 69–72

Roles for intelligence in multimedia: report on the IMMI-1 workshop

JOHN LEE

Human Communication Research Centre, University of Edinburgh, Edinburgh, Scotland

1 Introduction

Very little in computing these days is promoted with as much vigour as multimedia. Multimedia, we are told, will increase the usability and productivity of systems; it will improve the ability of people to learn from educational applications; it will combine naturally with other interface technologies to create a more natural interaction than ever before.

Most of what we see, however, falls well short of these expectations. Most multimedia systems are little more than a means of stringing together pre-packaged information which, almost incidentally, may take the form of images, sound or video clips, as well as simple text. The structure behind the presentation is usually limited to a hypertext network, where video clips, etc., can be the items at the nodes (which we are now used to calling "hypermedia"). Such systems are inevitably limited in their ability to respond to the needs of the user; they are often complicated and expensive to create; they are also complicated and expensive to change once created. To make good on the promises of the hype—if it is possible at all—we will have to develop systems that are much more responsive and adaptable (automatically or otherwise).

The objective of the First International Workshop on Intelligence and Multimodality in Multimedia Interfaces (IMMI-1), held in Edinburgh in July 1995, was to examine the prospects for such developments and identify pointers for promising research, given the present state of the art. For multimedia systems and their interfaces to become more responsive to the needs of users, they will have to have more "intelligence" in the sense of flexibility: for instance, they will have to be equipped with a more abstracted representation of the materials they are presenting. Where a current system might know that a node contains a bitmap of a certain size, what is often needed is some characterisation of what that bitmap represents an image of. Where a system may contain information (e.g., instructions for some task) in textual form, it may be that a user will benefit more from a graphical presentation, or even a short animation.

Problems like this call for a better understanding of "multimodality". Where "multimedia" tends to refer to presentation, "multimodality" has come to refer to interpretation and regeneration of information presented in different media. Construction of multimodal systems may call for advanced AI techniques including natural language and image processing. As a consequence, much in the area is experimental, or still at a mainly theoretical stage; but it may well be here that the future lies.

As usual, progress to date on such "intelligent" systems is much more tangible on the output side than in handling input. There have been significant advances in development of techniques for presentation of various kinds of data using a range of types of graphics, sometimes related to text, and even with some support for interactive dialogue, where for instance natural language text input is combined with pointing and clicking on a graphical display. Major further advances with this type of system are currently difficult because of a range of issues relating on the one hand to the desired input modalities—speech, gesture, drawings or sketches—and on the other hand to integration with various types of task and, ultimately, the old problem of knowledge representation. It was therefore natural, and indeed was part of the original conception of the workshop, that in many respects the outcome was yet more questions, rather than answers.

IMMI-1 raised these questions through a series of sessions in each of which about four presented papers were followed by a discussion open to the 60 or so participants. Since there were 27 presentations, the present account will not seek to summarise any of them in any detail, but rather to pick out themes that emerged as important. Further information about the workshop, the extended abstracts themselves, and hence also the names of co-authors, of whom none are mentioned here, are all available from the IMMI-1 World Wide Web pages, at http://www.cogsci.ed.ac.uk/~john/IMMI.)

The sessions were approximately focused around specific topics: language and learning, speech and retrieval, intelligence in multimodal presentation, software and interaction design, the relevance of formal theory, cognitive design and evaluation. In practice, the focus was usually a little diffuse because the workshop was characterised by an enormous diversity of papers addressing a wide range of different topics and approaches in the area. Thus among the presentations in the first two sessions, most ranged widely over the issues of language, speech, learning and retrieval.

2 Language and other modalities

An issue that became prominent in these first sessions was the synchronisation of communication in different modalities. If, for instance, language is to be combined with graphics—either the production of graphics, the animation of selected items, or the use of gesture by the user—then the establishment of co-reference relations between referential elements in each modality can be a significant problem. Nadia Bellalem described a remarkably general analysis of referential gestures in which the referring elements are isolated, typically as singularities in the shape and velocity of the curve traced by e.g. the mouse cursor or data-glove. In Jacques Siroux's GEORAL system, speech input is combined with gestures that may be simple pointing or the drawing of curves to designate a region forming the topic of an inquiry. In both these kinds of cases, there is a separation between the syntactic identification of the referential element in the gesture (analogous to, for example, a noun phrase in an utterance) and the analysis that yields its intended referent. This analysis usually can proceed only in a rich contextual environment which includes the interpretation of accompanying language; and often it turns out that the resolution of ambiguities depends on fine details of the relationship to this context. In practice, of course, compromises are made, and users often seem to adapt to them. This, however, leads to another kind of question: to what extent are we interested in supporting *natural* interaction, as opposed to *effective* interaction, where the latter is defined in relation to some task (or, for example, a learning outcome)? How far do these coincide?

John Dowell presented an approach to assigning information to modalities on the basis of a psychological model of "cognitive compatibility", developed in relation to an ambitious speech/graphics system called MASK (currently in use in French railway stations). Although it was possible on this basis to make some rather general recommendations, it had to be admitted in discussion that "the devil's in the details", and that users can be quite capricious in their responses to different styles of presentation, hardly ever behaving as one might have expected. This could range from the fact that MASK seemed only to attract a subgroup of possible users (mainly young computer programmers, it was said!), to the fact noted by several discussants that users of multimodal systems, where free choice is allowed, commonly get "hung up" on a particular modality, and appear reluctant to change to a different one even at times when it would be clearly more effective to do so.

Much of this discussion tended to gravitate towards problems of interaction—perhaps indicative of the high proportion of HCI-oriented participants. However, retrieval of multimodal information was recognised as a major issue in presentations by Jonathan Foote, on retrieving "video mail" by spotting words in messages, and Matt Hare, who focused on how to present the results of diverse forms of information found by searches in heterogeneous databases. Art Huntley described a range of approaches to using multimedia in medical education over the Internet, while Gavin

Long focused on the educational possibilities of natural language for dialogues in the context of a hypertext system, and Sergio Santana presented a system that integrates natural language and graphics in explanations of the solutions provided by a constraint-maintaining design drawing tool.

3 Theoretical approaches

Attempts to grapple theoretically with the complex diversity of multimodal phenomena take a number of forms. In the workshop, there were presentations ranging from attempts to taxonomise modalities, to formal theories aiming to capture their structure and to some extent their semantics. Thus Neils Ole Bernsen described an ambitious conception of a "modality theory" in which all possible combinations of the most basic modalities could be elaborated and related to their communicational possibilities. Constantine Karagiannidis, Mark Greaves, Fabio Paterno, Henk Zeevat and Jean-Claude Martin in different ways all also proposed theories of the logical content or combinational possibilities of modalities, sometimes very directly and sometimes (e.g., Greaves) via representational "interlingua". Winfried Graf focused on the sometimes overlooked issue of layout in text/graphics systems, which materially affects the relationship between these modalities while not being clearly assimilable to either. The problem of how to reason about or with multimodally presented information lurked constantly below this discussion, wriggling through various answers to the old Catch-22 problem that if multiple representations are commensurable enough to be reasoned across, they easily appear to be redundant.

All of these approaches can only be described at present as preliminary—the area is comparatively new and there is no well-established paradigm. All are necessarily limited in many respects to a particular point of view. The most salient point of discussion in this connection was the issue of how useful any of these theories can be in abstraction from some notion of the *task* the user is undertaking. This point has two aspects: on the one hand, we have no general grasp on the ways in which (task) context affects the actual use and interpretation of modalities, or the felicity of various ways of combining them; and on the other hand, it's often difficult to relate a formal account to the informal specifications and descriptions provided by users. This is partly just the usual difficulty of relating formal accounts e.g. of natural language semantics or proof theory to everyday experience of communication and reasoning; but it is exacerbated by the lack of accepted formal semantics for most modalities. Jean-Claude Martin, by contrast, proposed an emphasis on *types of cooperation* rather than *choices of media,* but it remains unclear how a level of analysis abstracted that far above the demands of a particular task will turn out to be applicable when considering specific cases. We are at a very early stage in addressing these issues, but the workshop at least showed that there are promising developments.

4 Cognitive approaches

What remains clear is that any theory which has reasonably general application is going to have to embrace not only the formal characterisation of modalities as information-carriers, but also the cognitive capacities of users as extractors, processors and providers of information. It is in the trade-off between formal and cognitive properties of representations that we will find out what is really redundant and what isn't. A fair range of the topics addressed by the workshop were related to these cognitive issues, from the two closely related angles of design and evaluation. As with the case of formal theories, there is in the area of multimedia little solidly agreed common ground of cognitive theory, hence the process of design and evaluation based on such theory often has as much the role of testing the theory as of producing a workable system. Cognitive theory can also be closely linked to a formal approach, especially if, for example, a task involving reasoning is in view. Keith Stenning, for instance, characterised the cognitive differences between language and graphics via a view of them as systems with differing expressive power. Without conventions that make it more "symbolic", graphics is a relatively inexpressive medium, and this accounts, for

instance, for the relative tractability of graphical representations for restricted classes of inferences.

Yvonne Rogers exploited aspects of this idea in attempting to address a wider range of issues about the evaluation of multimedia presentations, agreeing with Richard Cox in emphasising the central importance, when selecting external representations, of matching the information to be communicated to the expressive properties of the media or modalities used. Peter Faraday attempted to relate external to internal representation in a presentation of a new "walk-through" evaluation technique based on a mental-model theory. Kristina Höök emphasised the need for users to be able to understand the workings of a system, which should appear to them as a "glass box": she described an imaginative multimedia help system using both natural language and graphics. A cautionary note for designers of all such systems was sounded by Jon Oberlander's presentation on individual differences, in which he showed that the value of various kinds of multimedia representation, or the use of spatial metaphors (e.g. in hypertext navigation) can have very different effects depending on the spatial-reasoning ability of the users. Whereas the performance of one group of users may be substantially improved, others may actually be significantly hindered. This is clearly an area of key concern for research both in HCI and in education.

Given the emphasis, at various points in the workshop, on affective and motivational factors that can be associated with a task and significantly alter users' reactions, there was surprisingly little reaction to Detlev Zimmerman's interesting discussion of a system that generates synchronised music for a multimedia presentation by relating the intentions of the author to the "rhetorical" functions of various musical phenomena. This could well be a useful technique for short displays in situations where it makes sense to generate the whole presentation automatically.

5 Architectures and implementation

Last but not least, we need to consider software and interaction design for multimedia. In a session on this topic, Laurence Nigay presented aspects of the PAC-AMODEUS "melting-pot" mechanism for "fusion" of multimodal information. This is relatively domain-independent, but it was observed during the discussion that it does not account for logical relations in information. The notion of an "intelligent agent" is one often encountered: Doug Riecken discussed a system called "M", which is an architecture of communicating agents; but at the same time insisted that the term "agent" is widely abused and should be replaced with more domain-specific terminology in particular contexts. To some extent, the level of abstraction of "M" was mirrored in Hans-Werner Gellersen's proposal to define interaction styles independently from realisation in any particular modality—contrary to the usual practice in e.g. object-oriented systems. But however characterised, multimodal interaction will have to make use of new techniques, and new devices, such as the "powered trackball" described by Reinder Haakma, which provides force-feedback and is under evaluation within a very general layered feedback model. The fact is that the diversity of multimodal systems is reflected in a wide plurality of approaches to implementation, and there is as yet nothing like a unifying "reference model" or standard architecture—if such a thing is desirable.

6 Conclusion

Overall, then, the IMMI-1 workshop was nothing if not diverse. This had the disadvantage, perhaps, that discussion was rarely in great detail; but it had the advantage that issues could be seen to apply in closely related ways across a large range of different approaches. The overall impression was thus that while "intelligent multimedia" is a very infant area of study, there is much scope for convergence and mutual support in the emerging community of researchers involved. Multimodality will undoubtedly continue to be a "growth industry".