Generating Multimedia Briefings: Language Generation in a Coordinated Multimedia Environment

Kathleen R. McKeown*
Dept. of Computer Science
Columbia University
New York, NY 10027, USA
kathy©cs.Columbia.edu

Abstract

Communication can be more effective when several media (such as text, speech, or graphics) are integrated and coordinated to present information. This changes the nature of media specific generation (e.g., language generation) which must take into account the multimedia context in which it occurs. In this paper, 1 will present work on coordinating and integrating speech, text, static and animated 3D graphics, and stored images, as part of several systems we have developed at Columbia University. A particular focus of our work has been on the generation of presentations that *brief* a user on information of interest.

1 Introduction

In many contexts, explanations can be more effective if they use multiple media. In our work, we have explored the generation of multimedia explanations in the context of providing instructions for equipment maintenance and repair [Feiner and McKeown, 1991], providing briefings on patient status after bypass for various caregivers {Dalai *et* a/., 1996a; 1996b], and providing illustrated briefings over online documents in an Internet environment [Aho *et* a/., 1997]. A key characteristic of our work is the dynamic generation of content and form at the time an explanation is required. This means that we can vary not only what is communicated, but also how different media are used in combination to best convey information depending upon the situation.

Our work has integrated a range of media within an explanation, including the use of written text accompanied by static graphics, the use of spoken language, text, and animated graphics, and the use of textual summaries with representative images. Our research has focused on coordinating the different media in a single explanation, using individual media to present aspects of the presentation for which they are particularly suited and ensuring that generation within one medium enhances portions of the presentation in other media.

The nature of generation for an individual medium, such as language, is changed by the fact that it occurs in a multimedia environment. For example, how language refers to information and entities in the domain changes when there is an accompanying illustration of the same information or entities. Furthermore, the nature of this change depends upon the specific media used. For example, coordination of spoken language and animated graphics requires that spoken references be temporally synchronized with the accompanying graphical reference. The combined use of text and speech, on the other hand, changes the content of generated references. Spoken references can be shorter and more natural as long as the accompanying textual reference provides the full, unambiguous reference.

In our more recent work, multimedia explanations are generated for the goal of *briefing* a user [Dalai *et* a/., 1996a; Aho *et* a/., 1997]. In the healthcare domain, they provide a concise summary for time-pressured caregivers. In the Internet environment, they highlight information contained in the underlying multimedia documents. In both scenarios, the nature of language generation differs from traditional text generation: summaries must be concise, using as few words as possible; at the same time, they must be informative, conveying as much information as possible in limited time or space.

In this paper, we present issues for language generation when produced as part of a multimedia briefing. We begin with a brief overview of the approaches we use in the three systems we have developed. We then discuss the modification of language generation for the multimedia context and finally, highlight questions for summarization.

2 An Overview of Columbia University Multimedia Systems

In this section, we provide a brief overview of several multimedia systems developed at Columbia University, highlighting domain, types of media used, and the goals for generating explanations. For full details on each system, refer to the original papers cited below, from which

*A great number of people have collaborated on the projects that are overviewed in this abstract of my invited talk. See the Acknowledgments Section for a list of contributors who in other situations have been co-authors.

the examples of generated explanations are drawn.

2.1 COMET

COMET (Coordinated Multimedia Explanation Testbed) generates coordinated, interactive explanations that combine text and three-dimensional graphics [Feiner and McKeown, 1991]. COMET not only determines what information to communicate using a *content planner,* but also how to express it in each medium using *medium-specific generators* [McKeown *et ai,* 1990; Seligmann and Feiner, 1991; McKeown *et ai,* 1993], Text and graphics are coordinated by communication with a *media coordinator* [Feiner and McKeown, 1990; McKeown *et al,* 1992]. COMET was developed to provide explanations for equipment maintenance and repair. It generates explanations that instruct users how to carry out diagnostic tests on a particular piece of equipment, a military radio receiver-transmitter. Explanations typically describe one or more steps in these tests that are presented in a series of displays, where each display includes an illustration and a caption providing the textual instruction.

2.2 MAGIC

MAGIC (Multimedia Abstract Generation for Intensive Care) is being developed to provide a multimedia interface to health care data. In particular, MAGIC is designed to provide briefings on patient status immediately following a coronary bypass operation. In a Cardiac Intensive Care Unit, communication regarding patient status is critical during the hour immediately following bypass. It is at this critical point, when care is being transferred from the operating room to the Intensive Care Unit and monitoring is at a minimum, that the patient is most vulnerable to delays in treatment. During this time, there are a number of caregivers who need information about patient status and plans for care. Yet, the only people who can provide this information are those who were present during surgery and they are often too busy attending to the patient to communicate much detail.

MAGIC takes as input online data collected during the surgical operation as well as information stored in the mainframe databases at Columbia Presbyterian Medical Center [Roderer and Clayton, 1992]. It generates a multimedia briefing that integrates speech, text, and animated graphics to provide an update on patient status [Dalai *et ai,* 1996a]. Like COMET, it dynamically determines both content and form of the explanation, but the focus here is on the coordination of temporal media [Dalai *et ai,* 1996b]. Language generation addresses the issue of producing language (wording and sentence structure) appropriate for the spoken medium [Pan and McKeown, 1996; McKeown *et ai,* 1997] as opposed to the more traditional task of generating written language. Given the healthcare setting, MAGIC also faces the added constraint of conciseness; the generation process must make coordinated use of speech and text to produce an overview that is short enough for time pressured caregivers to follow, but unambiguous in meaning.

2.3 CDNS

Research in CDNS (Columbia Digital News System) focuses on the development of technologies to aid people in finding and tracking information on current events. We are developing a system, CDNS, that provides up-to-the-minute briefings on news of interest, linking the user into an integrated collection of related multimedia documents. Our research aims at tracking news stories on the same event, producing a briefing that describes how the event has changed over time. A representative set of images or videos can be incorporated into the summary. The user can follow up with multimedia queries to obtain more details and further information.

The goal of summarization within CDNS is to brief the user on information within the collection of related documents. In many cases, the summary provides enough information for the user to avoid reading the original document. In others, the user may want to check the original documents to verify information contained within the summary, to follow up on an item of interest, or to resolve conflicting information between sources. To meet the demands of this environment, our work features summarization over *multiple articles,* merging information from all relevant sources into a concise statement of the facts. This is in contrast to most previous work that summarizes single articles [Luhn, 1958; Paice and Jones, 1993; Rau *et ai,* 1994; Kupiec *et ai,* 1995]. Summaries must identify how perception of the event changes over time, distinguishing between *accounts* of the event and the event itself [McKeown and Radev, 1995]. Eventually, given access to live news, the summarizer will be able to provide updates since the last generated summary, identifying new information and linking its presentation to earlier summaries.

Unlike COMET and MAGIC, CDNS must determine the content of a briefing through analysis of input textual documents. Furthermore, while CDNS generates the content and form of the written summary, images are selected as a whole from an online collection of images.

3 Language Generation in a Multimedia Context

In a multimedia explanation each medium presents information about the same subject. In some cases, the exact same information is presented in multiple media, providing different viewpoints of the same material. In other cases, media may present complementary information about the same topic when certain types of information are more easily communicated in one medium than another. For example, spatial information such as location may be more easily conveyed in graphics, while abstract information such as illness severity may be more easily conveyed in language.

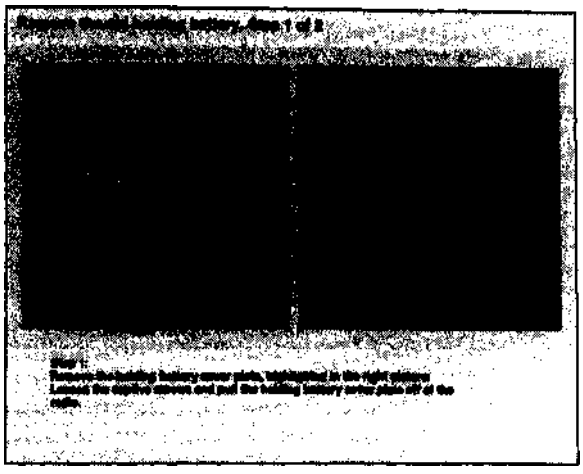

Figure 1: Cross reference generation in COMET

The fact that each medium presents information about the same topic, sharing communicative goals, means that individual media can not be generated in isolation. How individual media generation is affected by other media depends on the types of media that are coordinated within the multimedia presentation. A multimedia presentation that coordinates speech and graphics engenders different influences on the language generation process than one that coordinates text and images.

As an example of the kind of influences that exist, consider that at the most basic level, different media refer to the same information or entities within the same multimedia presentation, often to satisfy the same high level goal. If they have knowledge about the references used in other media, they can use that to influence how referring expressions are generated in their own medium. In addition to changing how referring expressions are generated, in some cases an explicit cross reference may be effective to link information in one media with another. Exactly how the generation of referring expressions is modified varies depending upon the media involved.

Text and static graphics. In our earlier work on COMET, multimedia presentations comprised written language and static graphics. In this scenario, text appears as a caption below an illustration. To signal where text and illustration refer to the same object or action, an explicit cross reference must be generated linking the two media. Here, generating a textual cross reference must make use of knowledge about the graphical illustration to determine what kind of cross reference to generate. The language generator can make use of information about the illustration content to identify a referent for the hearer; the text "the old holding battery, *shown in the cutaway view"* uses information about special graphical features used in the illustration (here graphics cut away a portion of the radio to reveal the holding battery which is internal to the radio). Alternatively, it can use information about illustration structure as shown in Figure 1 or information about spatial location. To generate cross references, language generation must be modified to determine when a cross reference is needed (e.g., when a user will not know the referent of a textual expression alone) and query the illustration representation to construct cross reference content. In COMET, generating cross references is carried out cooperatively by the graphics and text components and can involve modifications to the illustration as well as to language [McKeown *et* a/., 1992].

Spoken language and animated graphics. When spoken language and animated graphics are part of the multimedia presentation as is the case in MAGIC, a more implicit means of linking references across media can be used. As references are spoken, graphical representations of the referenced information can be highlighted in the accompanying illustration. In this scenario, spoken references must be coordinated with accompanying highlighting that changes over time, involving negotiation between speech and graphics to arrive at both a compatible ordering and duration of references [Dalai *et* al., 1996b]. In particular, speech has grammatical constraints on how references are linearly, and thus temporally, ordered within a sentence. At the same time, graphics has spatial constraints on how information is arranged within an illustration. Taken individually, such constraints might result in an incoherent presentation which refers to graphical representations at random locations in the illustration. Similarly, durations of both spoken references and highlighting must be coordinated to avoid changes in highlighting at semantically anomalous points in the sentence or blinking that might occur if highlighting is changed too frequently.

The problem for language generation in this context is to produce enough information to facilitate coordination. First, note that the full ordering of spoken references is only determined when all grammatical constraints have been applied and the final sentence generated. But it would be quite inefficient to wait until this point to coordinate with graphics as it could potentially involve generating many sentences that were never used in the final presentation. Instead, we produce a partial ordering over references at an intermediate point in language generation, after words have been selected but before grammatical constraints are applied. Second, given that any particular ordering of spoken references produced by the language generation component may not be compatible with graphics' orderings, the language generation process must be modified to produce several possibilities ordered by preference. Third, since graphics does not understand the meaning of a string of words that forms the spoken reference, the language generator must maintain a mapping between the words of the reference and the semantic object they refer to in order to communicate with graphics via the media coordinator. Finally, to facilitate synchronization of spoken references with highlighting, the spoken language generator must be able to compute duration of a spoken reference, to reason about

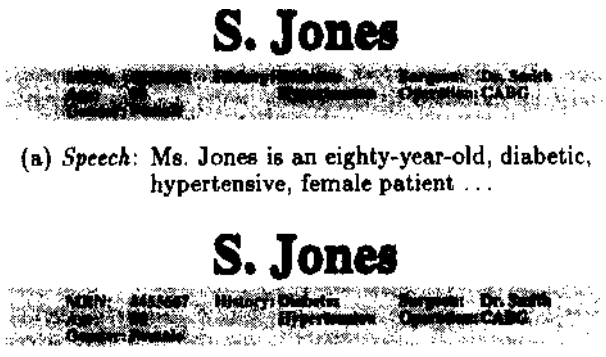

Figure 2: A portion of coordinated speech and graphics generated by MAGIC

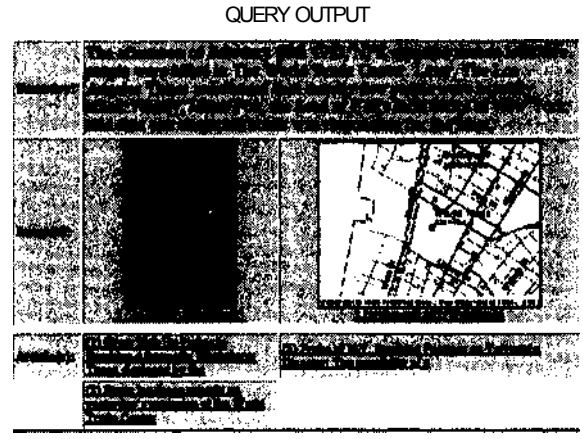

Figure 3: Illustrated summary generated by CDNS

where pauses can be adjusted in speech to allow more flexibility in coordinating with highlighting, and to select semantically appropriate points in the sentence between references which can be temporally coordinated with changes in highlighting [Pan and McKeown, 1996].

Figure 2 shows a portion of a multimedia briefing generated by MAGIC. Here, as each part of the first sentence is spoken, the corresponding information in the demographics chart of the accompanying illustration is highlighted. Just the chart is shown in Figure 2.

Spoken and written language. In MAGIC, both speech and text are used within the same presentation. Textual references are used to provide labels for objects and information displayed graphically. Language generation takes advantage of the use of both media to keep spoken language shorter and more colloquial, thus better meeting our goal of briefing caregivers. As long as the text label on the screen is generated using the full, unambiguous reference, speech can use an abbreviated expression. For example, when referring to the medical devices which have been implanted as part of cardiac care, speech can use the term "pacemaker" so long as the textual label specifies it as "ventricular pacemaker". Similarly, MAGIC uses "balloon pump" in speech instead of "intra-aortic balloon pump", which is already shown on the screen. In order to do this, lexical choice in both media must be coordinated. Lexical choice for text always selects the full reference, but lexical choice for speech must check what expression the text generator is using. The speech lexical chooser must check what attributes the text generator includes in its reference and omit those.

Written language and images. In CDNS, a written summary is generated along with several representative images as shown in Figure 3. In this scenario, the specific objects contained in the image are unknown[1], and therefore, the written summary cannot explicitly refer to image content. Unlike the other multimedia systems discussed here, the content and form of the accompanying image is not dynamically generated at the time of the presentation. Instead, an image is selected for use when the briefing is generated. The problem here is to ensure that the images selected are relevant to the content of the summary. We are exploring the use of integrated processing of textual and image features to select appropriate images. This is done by classifying unlabeled images as part of an underlying ontology. Image features can be used to reliably make certain categorizations; for example, image features can be used to determine if the image is a graphic, such as a map, or a photograph as well as more semantic categorizations such as whether it is a portrait showing a person or a landscape [Smith and Chang, 1996]. We are augmenting this by using statistical analysis of the text accompanying an image to provide a better semantic classification of image content (e.g., whether an image pertains to a terrorist event or a particular political event such as a Russia-US summit) [Aho *et* al., 1997]. By applying standard text categorization techniques to different amounts of text surrounding an image in a multimedia document, from caption to text fragment to full document, we are experimenting with the improvement on classification over using image features alone.

4 Issues for Summarization

Just as the nature of language generation changes in a multimedia context, so do problems for summary generation change from the generic problems in language generation. Summaries must convey maximal information in a minimal amount of space. This requires selecting words and sentence structures that can convey information concisely. Sometimes, this means the

[1] Unless we use vision techniques to do image analysis and identify the objects the image contains, a prospect that is not yet feasible for domain independent image processing, we cannot reliably identify and label image content.

use of complex sentence structure, including multiple modifiers of a noun or verb, conjunction (e.g., "and"), and ellipsis (i.e., deletion of repetitions across conjoined phrases). For example, phrases such as "hypertensive" and "undergoing CABG" use fewer words than if full sentences were used to convey each of these facts separately (see Figure 2). Conciseness also means the selection of words that can convey multiple aspects of the information to be communicated. For example, a verb such as "surged" conveys both the direction and the speed of a gain on the stock market. Furthermore, our research shows that some information is opportunistically added into the summary, depending on the words and syntactic structure already used [McKeown *et* al., 1995; Robin and McKeown, 1996].

We characterize problems for summary generation as falling into two separate classes, *conceptual summarization,* or determining what information should be included in a summary, and *linguistic summarization,* the task of determining how to convey as much information as possible in a short amount of text. Conceptual summarization takes input from multiple sources, whether databases or text, and determines how it can be merged together, often using semantic generalization to do so. Our work in CDNS addresses problems in conceptual summarization. Information is extracted from each article and represented in a template using systems developed under the DARPA message understanding program [MUC, 1992]. CDNS then uses planning operators to determine how to merge information from the separate templates representing each article. In particular, it looks for contradictions, agreements, and refinements of information, and makes generalizations.

Our work in MAGIC addresses problems in linguistic summarization. As in previous work on summarization over data [Robin and McKeown, 1993; 1996], we address the following issues:

- How to use syntactic and lexical devices to convey information concisely?
- Given the choice of a particular word or syntactic structure, how does this constrain (or allow) the attachment of additional information?
- How to fold multiple pieces of information into a single linguistic construction?

In MAGIC, this means conveying as many attributes from the underlying database as possible in one sentence through the use of modifiers such as adjectives or prepositional phrases. Thus, in Figure 2, the first sentence conveys nine separate attributes. In addition, coordinated use of speech and written language also aids in meeting our goal of conciseness; shorter references can be generated in speech since they are clarified in the written labels of the accompanying illustration.

5 Conclusion

In this extended abstract, we have outlined issues for the generation of multimedia briefings and shown how we have addressed them in several different systems. While in this paper we focused on the problem of generating references in a multimedia environment, we have also looked at the interaction between generation in different media for other problems as well. For example, in work on COMET we explored how separation of information into different pictures can influence sentence breaks. In general, style of generation in one media can and should influence generation in others.

Of course, there are many other issues in the generation of multimedia briefings. In our current work, we are addressing the generation of different types of prosody for speech using information from language generation. We are particularly interested in the use of prosody that facilitates coordination within the multimedia environment such as pause duration. For example, by computing a range on pause duration and options on pause placement, we simplify the task of temporally synchronizing spoken and graphical actions. In CDNS, we are continuing to exploit multimedia features in online news sources to improve both generation of illustrated briefings and search over online, multimedia documents. We are also working on the generation of summaries from live information that update a user over information already received.

Acknowledgments

This extended abstract reflects the work of many individuals over the course of many years. My closest collaborator in work on multimedia explanation has been Steven Feiner, with whom I have enjoyed many long hours of thought-provoking discussion, joint meetings on system design and development, and mad races to deadlines. In our recent work on MAGIC, Mukesh Dalai has also joined our collaboration on multimedia coordination. Others who have contributed to work described here include: Michael Elhadad, Doree Seligmann, and Jacques Robin (COMET); Desmond Jordan, Barry Allen, Shimei Pan, James Shaw, Michelle Zhou, Tobias H611erer, and Yi Lang (MAGIC); Alfred Aho, Shih Fu Chang, Dragomir Radev, John Smith, Alex Jaimes, and Kazi Zamen (CDNS).

References

[Aho *et* a/., 1997] A. Aho, S.-F. Chang, K. R. McKeown, D. Radev, J. Smith, and K. Zaman. Columbia Digital News System: An environment for briefing and search over multimedia information. In *Proceedings of ADL-97,* Washington, D.C., May 1997.

[Dalai *et ai,* 1996a] M. Dalai, S. Feiner, K. R. McKeown, D. Jordan, B. Allen, and Y. alSafadi. MAGIC: An experimental system for generating multimedia briefings about post-bypass patient status. In *Proc. 1996 AMIA Annual Fall Symp.,* pages 684-688, Washington, DC, October 26-30, 1996.

[Dalai *et ai,* 1996b] M. Dalai, S. Feiner, K. R. McKeown, S. Pan, M. Zhou, T. Hoellerer, J. Shaw, Y. Feng,

and J. Fromer. Negotiation for automated generation of temporal multimedia presentations. In *Proc. ACM Multimedia '96,* pages 55-64, Boston, MA, November 18-22, 1996.

[Feiner and McKeown, 1990] S. K. Feiner and K. R. McKeown. Coordinating text and graphics in explanation generation. In *Proceedings of the National Conference on Artificial Intelligence,* Boston, Mass., August 1990.

[Feiner and McKeown, 1991] S. Feiner and K. R. McKeown. Automating the generation of coordinated multimedia explanations. *IEEE Computer,* 24(10):33-41, October 1991.

[Kupiec *et al,* 1995] Julian M. Kupiec, Jan Pedersen, and Francine Chen. A trainable document summarizer. In Edward A. Fox, Peter Ingwersen, and Raya Fidel, editors, *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* pages 68-73, Seattle, Washington, July 1995.

[Luhn, 1958] Hans P. Luhn. The automatic creation of literature abstracts. *IBM Journal,* pages 159-165, 1958.

[McKeown and Radev, 1995] K. R. McKeown and D.R. Radev. Generating summaries of multiple news articles. In Edward A. Fox, Peter Ingwersen, and Raya Fidel, editors, *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* pages 74-82, Seattle, Washington, July 1995.

[McKeown *et ai,* 1990] K. R. McKeown, M. Elhadad, Y. Fukumoto, J. Lim, C. Lombardi, J. Robin, and F. Smadja. Language generation in COMET. In *Current Research in Language Generation.* Academic Press, London, 1990.

[McKeown *et al,* 1992] K. R. McKeown, S. K. Feiner, J. Robin, D. Seligmann, and M. Tanenblatt. Generating cross references for multimedia explanations. In *Proceedings of AAAI-92.* AAAI, July 1992.

[McKeown *et al,* 1993] K. R. McKeown, J. Robin, and M. Tanenblatt. Tailoring lexical choice to the user's vocabulary in multimedia explanation generation. In *Proceedings of the 81st Annual Meeting of the Association for Computational Linguistics,* Columbus, Oh., June 1993.

[McKeown *et al,* 1995] K. R. McKeown, J. Robin, and K. Kukich. Generating concise natural language summaries. *Information Processing and Management, special issue on summarization,* 31(5):703-733, September 1995.

[McKeown *et al,* 1997] K. R. McKeown, S. Pan, J. Shaw, D. Jordan, and B.A. Allen. Language generation for multimedia healthcare briefings. In *Proceedings of the ACL Conference on Applied Natural Language,* Washington, D.C., April 1997.

[MUC, 1992] DARPA Software and Intelligent Systems Technology Office. *Proceedings of the Fourth Message Understanding Conference (MUC-4),* 1992.

[Paice and Jones, 1993] Chris D. Paice and Paul A. Jones. The identification of important concepts in highly structured technical papers. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* pages 69-78, 1993.

[Pan and McKeown, 1996] S. Pan and K. R. McKeown. Spoken language generation in a multimedia system. In *Proceedings of ICSLP 96,* volume 1, pages 374-377, Philadelphia, PA, 1996.

[Rau *et al,* 1994] L. F. Rau, R. Brandow, and K. Mitze. Domain-independent summarization of news. In *Summarizing Text for Intelligent Communication,* pages 71-75, Dagstuhl, Germany, 1994.

[Robin and McKeown, 1993] J. Robin and K. R. McKeown. Corpus analysis for revision-based generation of complex sentences. In *Proceedings of the 11th National Conference on Artificial Intelligence,* Washington, D.C., July 1993. American Association for Artificial Intelligence.

[Robin and McKeown, 1996] J. Robin and K. R. McKeown. Empirically designing and evaluating a new revision-based model for summary generation. *Artificial Intelligence,* 85, August 1996. Special Issue on Empirical Methods.

[Roderer and Clayton, 1992] N. Roderer and P. Clayton. IAIMS at Columbia Presbyterian Medical Center: Accomplishments and challenges. In *Bull Am. Med. Lib. Assoc,* pages 253-262, 1992.

[Seligmann and Feiner, 1991] D. Seligmann and S. Feiner. Automated generation of intent-based 3D illustrations. In *Proc. ACM S1GGRAPH '91* ("Computer Graphics, *25(4), July 1991),* pages 123-132, Las Vegas, NV, July 28-August 2 1991.

[Smith and Chang, 1996] J.R. Smith and S.-F. Chang. Searching for images and videos on the World-Wide Web. Submitted to IEEE Multimedia Magazine, also CU-CTR Technical Report 459-96-25, 1996. Demo accessible from URL http://www.ctr.Columbia.edu/webseek.