Introduction

Mark T. Maybury
The MITRE Corporation
202 Burlington Road
Bedford, MA 01730
maybury@mitre.org

Abstract

Our lives are increasingly surrounded by data, information and knowledge captured in multiple media: text, graphics, imagery, audio, and video. These media are frequently combined or structured to form complex artifacts (e.g., hypermedia documents, interactive CD-ROMs) which exploit multiple perceptual modalities (e.g., auditory, visual, haptic/gestural). This book focuses on tools and techniques that support efficient and effective indexing, browsing, retrieval, interaction with and visualization of multimedia. Multimedia digital libraries which incorporate text, graphics, audio, and video are central to many applications areas including information access, training, and decision support. This chapter introduces the need for intelligent multimedia information retrieval, outlines its theoretical foundations, outlines the current state of the art, describes the structure of this collection, and outlines some remaining fundamental problems.

1. Purpose and Scope

Increasing use and expansion of the information highway has created requirements for new and improved access to global and corporate information repositories. These repositories increasingly go beyond free text and structured databases to include graphics, imagery, audio (speech, music, sound), and video artifacts. The advent of large, multimedia digital libraries has focused attention on the problem of enabling more efficient and effective multimedia information access.

Traditionally, largely independent research communities have focused on the automated processing of single media including text processing (Grosz, Sparck Jones and Webber 1986; MUC-6, 1995), spoken language processing (Waibel and Lee 1990), and image and video processing (Niblack and Jain 1993-95; Chen, Pau and Wang 1993; IFIP 1989, 1992; Furht, Smoliar and Zhang 1996). The challenge of massive, multimedia digital libraries has turned attention toward the problem of integrated access to structured data and textual sources as well as media with spatial and temporal properties (e.g., sound, maps, images, video), the focus of this book. This collection differs from previous works focused on the more general issues of human computer interaction (Baecker et al. 1995), multimedia or intelligent interfaces (Blatner and Dannenberg 1992, Sullivan and Tyler 1991), intelligent multimedia interfaces (Maybury 1993) or multimedia systems issues such as standards, compression/storage, communications and networking (Furht 1996). In contrast, this edited collection targets fundamental issues in processing and providing content-based, tailored access to multimedia artifacts (e.g., documents, video mail, and broadcasts), typically using intelligent or knowledge based techniques to do so. The reported investigations aim to create a broad spectrum of new capabilities for a range of media including multimedia information browsing, search, extraction, visualization, and summarization. Results of these endeavors promise new applications such as customized television, interactive radio, and content-based multimedia authoring tools.

As such, solutions to some of the fundamental problems in this area need to draw upon and integrate results from many disciplines. These include information retrieval, cognitive science, software design, human computer interaction, computer graphics, database management, and artificial intelligence and its subareas (e.g., vision, speech and language processing, knowledge representation and reasoning, machine learning/knowledge discovery, planning and agent-modeling). Only through a collaborative effort will we make the necessary advancements to more toward a more principled understanding of multimedia information processing.

2. Theoretical Foundations

The analysis of information includes several key processes: detecting relevant sources, translating/converting them, extracting information from them, and exploiting this information (see Figure 1). Detecting information could include identifying a relevant document as the result of a keyword search of several text databases, spotting keywords in a spoken language stream, or recognizing a face in a set of images. Having detected a relevant source, it may require translation from a source to a target natural language (e.g., English to Spanish text), conversion from one media to another (e.g., speech-to-text spoken language transcription), or mapping from a statistical representation to a symbolic one. Having the information in a common form, content (i.e., objects and their properties) can then be extracted. Examples of information exploitation include browsing a document collection, visualizing retrieved documents to detect patterns, looking up specific extracted facts, summarizing extracted information, or further processing it to identify correlations and trends.

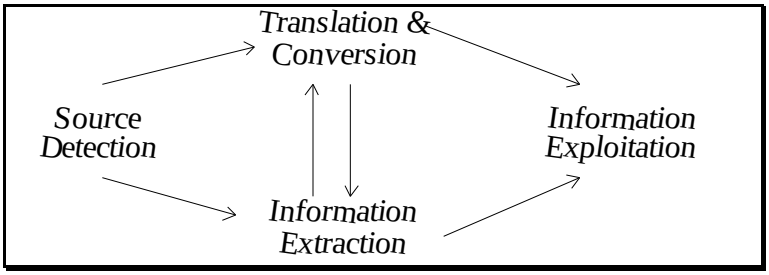

Figure 1. Information Analysis

Most research and commercial development in document detection has focused on text retrieval. In the traditional information retrieval model (Salton 1988, van Rijjsbergen 1979), an index (I) is constructed from a set of documents (D_i) to which queries (Q_i) are applied by users to satisfy some information need. Document indices can be constructed by a variety of methods including statistical means (e.g., creating histograms of letter or word frequencies in documents) and linguistic analysis (e.g., parsing and interpreting natural language found in the documents). As a consequence, indices found in commercial and research systems range from a simple inverted index (i.e., words or word-stems and pointers to their occurrence in D_i), to vectors of linguistic features found in documents (e.g., inferred document subjects) to structured databases of entities extracted from the source text from natural language processing (i.e., the who, what, when, where, why described in the documents). Queries can range from simple keywords to free form, natural language text. Query processing is often performed using the same techniques used for document indexing, although specialized processing is sometimes performed (e.g., term expansion using thesauri). Retrieval is the process of matching queries to documents. Evaluation of system effectiveness in satisfying user information

needs (as characterized by Q_i) is typically measured with large annotated corpora using *recall*, a measure of the ability of the system to retrieve all relevant documents in D_i and *precision*, a measure of the system's ability to return only relevant documents. In other terms, system precision is one minus the number of false positives; system recall is one minus the number of false negatives. Current areas of research include the use of language processing to deepen the level of document processing, scaling to massive document collections, dealing with heterogeneous collections, and foreign language document retrieval/information extraction.

Extensions of the above model are required to deal with documents in forms other than text and to deal with access across heterogeneous document collections. Following Maybury (1993), we define *medium* as material centered -- entailing both the physical media objects (e.g., ink on paper, soundwaves, video tape) as well as the logical means by which information is conveyed (e.g., natural language, sign language). The interpretation of media relies upon human centered processes, namely sensory *modalities* such as visual, auditory, and tactile perception as well as of course higher level cognition. Complementary production modalities in humans include: writing, speaking and gesture (to include hand, head, eye, and body motion). Consequently, a particular media (e.g., language) can be conveyed in multiple modalities (e.g., spoken or written language). *Multimedia information retrieval*, then, is the use of computer programs to access digital libraries of multiple media (e.g., text, audio, imagery, video). *Intelligent* multimedia information retrieval goes beyond traditional hypertext or hypermedia environments to provide content based indexing of multiple media and management of the interaction with these materials by representing and reasoning about models of the media, user, discourse and task.

As Figure 2 illustrates, principal multimedia information processes include:

- *Analysis* of the multimedia artifacts (e.g., text, graphics, video) to index them or extract information from them (e.g., the objects, relations, and events contained in or communicated via the media).

- *Retrieval* of indexed information from single and multiple media document collections using single and cross-media query languages to support viewing of, interaction with, or generation of multimedia documents.

- *Generation* (planning or realization) of new, possibly multimedia artifacts from existing repositories.

- *Interaction* with existing collections, drawing upon the above processes but also dealing with tailoring access to the user, task, and/or situation.

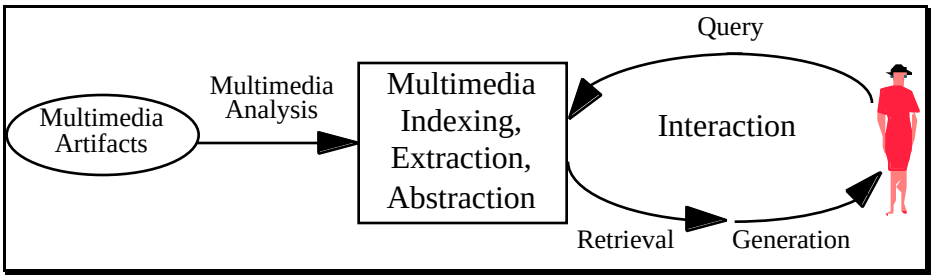

Figure 2. Multimedia Processes

Standard techniques for mapping between *Q*, *D*, and *I* are text-centric and need to be extended to address different media. For example, the document index, *I*, will need to be extended to include non-text media characteristics, such as the intonation and pitch contours of spoken language, the color, size, and shape characteristics in graphics and still imagery, and the camera-effects, lighting and motion captured in video. Creation of a multimedia index needs to address the fact that different media have different information carrying properties (e.g., static vs. dynamic, ease of representing quantitative vs. qualitative or temporal versus (geo)spatial information). This raises the need to mediate among media specific representations. For example, because of an inability to perform object and event recognition in all media, some indices will remain statistical (e.g., color histograms from images) whereas others will be symbolic (e.g., extracted names, organizations, and locations mentioned in text documents). Providing an integration of representations -- a *media interlingua* -- or at least access across heterogeneous media representations, remains an important challenge.

Figure 3 illustrates how media can be described and/or represented at multiple levels of abstraction, each increasingly higher level of which could be used for indexing for retrieval (up arrows) or presentation generation (down arrows). That is, as you proceed from the bottom to the top of the figure, which can be viewed as a set of transformations among representations, you reach successively deeper levels of representation, moving from surface forms to underlying intentions (if they exist). Each media has basic *elements* (which have associated attributes or features), a syntactic means of grouping, ordering and structuring these (e.g., a natural language grammar, a grammar of spatial or temporal constraints), and an associated semantic and/or conceptual meaning. At the highest level, an artifact in a media can serve an intentional purpose, e.g., a sentence can inform or request, a graphic can persuade, an image can shock or delight. While there remains debate regarding the basic unit of speech (e.g., phone, syllable), in text the word and its morphological variants and associated orthographic attributes (e.g., font, size, color) generally serve as basic elements. Words can be structured into (syntactic) phrases, utterances or sentences, and interpreted using compositional semantics. Analogously, in line graphics, visual elements such as pixel, lines, and regions have associated features (e.g., size, location, color) and can be organized into visual objects which in turn may encode some conceptual meaning (e.g., a person, location, or event), which in turn might perform some communicative action (e.g., inform, surprise).

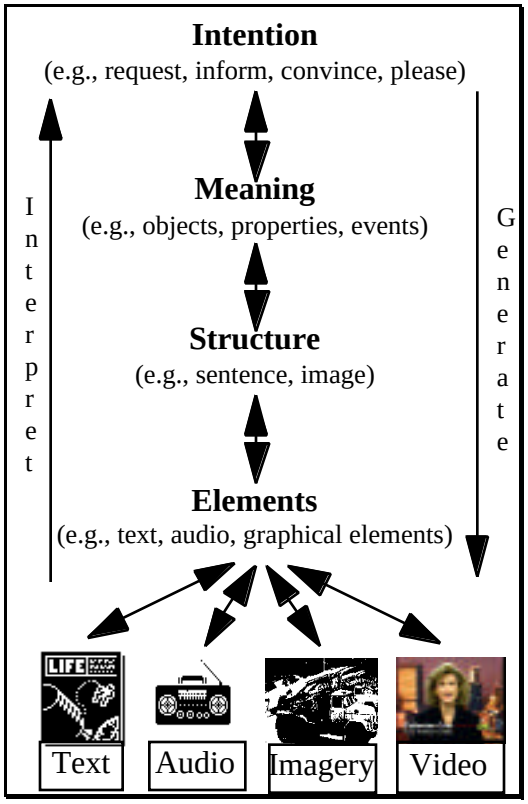

Figure 3. Levels of Representations for Media

Important relations can exist across media at all levels. For example, text labels on a map can be encoded using color coordinated with a graphic or image. A sound or image could be used to convey linguistic content (e.g., a handwritten image; a text scrolled in a video). Cross media references might call out images by using text or graphics with an audio overlay. The state of the art is characterized by varying degrees of ability to both explicitly represent and automatically process media at these various levels of abstraction. For example, whereas there exist a broad range of text processing methods for lexical, morphological, syntactic, semantic, and, to a more limited extent, discourse and pragmatic analysis, we do not have agreed upon methods of describing the basic elements of imagery and audio. For example, as Blum et al. (this volume) address automated indexing of sounds, features such as pitch, loudness, and duration are readily describable and computable, however, timbre (related to the amplitude envelope, harmonicity, and spectral envelope), characterizing for example the tonality of an instrument, is primarily a perceptual concept, and thus derived from acoustic properties. Blum et al. are thus led to develop both acoustic and perceptual indices of audio. Similarly, whereas image properties such as color and shape are directly computable, as noted in this volume by Flickner et al. and Zhang et al., more sophisticated notions such as texture do not have agreed upon definitions, rather they are associated with multiple properties, including the directionality, granularity or coarseness, and contrast of an image. Thus, it is to be expected that terminological refinements will occur as we begin to increase our understanding of respective media and their relations. Experience with higher levels of representation (e.g., discourse, pragmatics) in non-textual media (e.g., video, animation) will expand our previous text-centric notions to take on new dimensions. For example, traditional models of focus of attention and discourse (Grosz and Sidner 1986) will need to be extended to address spatial and temporal focus of attention (Maybury 1991) as well

as multimedia co-reference; models of communicative actions will need to incorporate visual and auditory actions; and user models will need to capture skills with and preferences for various media types and properties.

Document and query analysis techniques similarly need to be extended to enable, for example, users to posit queries not only by keyword and topical interest but also by color, shape, and texture for imagery; by gender, intonation, or rate of speaking in a spoken language stream; by characterizing the pitch, tone, and timbre for retrieval of sounds. Similarity based retrieval (i.e., "its like this image, sound bite, or video clip"), relevancy ranking of results, and results visualization will become important aspects of multimedia information access as the scope and complexity of the media space increases. When positing cross media queries (e.g., "find me all documents containing moving pictures of or spoken references to the President that have a duration of more than 2 minutes"), the multimedia index must support retrieval of non-text media via temporal, spatial, and visual properties.

3. State of the Art: An Overview of the Collection

This collection is organized around the following seven sections which, collectively, offer mechanisms to index, browse, search, visualize, and interact with imagery, graphics, audio, and video. The first five sections focus primarily on indexing multimedia, the sixth focuses on interacting with previously indexed or structured material, and the last focuses on empirical assessments of user performance in multimedia information spaces.

3.1 Content-based Access to Imagery

The chapters in the first section of the book focus on methods for automated indexing of imagery to support search and browsing. The first chapter by Flickner et al. describes the Query By Image Content (QBIC) system (now in IBM's DB2 Image Extenders), which indexes imagery and video on the basis of visual properties such as color, shape, and texture as well as motion analysis. In the second chapter, Smith and Chang describe indexing of imagery based on visual features, including color and spatial layout, reporting precision and recall experiments on 3,100 images that demonstrate the value of query by example. Manmatha and Croft next describe the use of image equivalence classes of scripted words together with affine transformations to support more effective retrieval of hand written historical manuscripts. In the last chapter, Rowe and Frew comparatively evaluate two techniques for image classification (case based and neural network) in a large, real world database of natural photographs. They show how associations between a "visual focus" computed from the image and a "linguistic focus" computed from the image captions can together improve classification performance.

3.2 Content-based Audio and Graphics Retrieval

The second section of the book addresses content based access to graphics and audio. In the first chapter, Roth et al. describe indexing of data-graphics to support indexing and retrieval based on graphical objects and relationships as well as structural similarity (e.g., find all graphs containing a descending line and an ascending one). Blum et al. then describe a user-extensible sound classification and retrieval system that computes both acoustic and perceptual properties to enable content based audio clip access.

3.3 Content-based Access to Video

In the first chapter in section three, Zhang et al. describe their video parsing algorithms for broadcast news which take advantage of models of anchor and reporter shots to go beyond transition detection (e.g., cuts, fades) to classify shot segments. The second chapter by Phillipe Aigraine presents a rule-based approach that infers the macrostructure of such videos as documentaries using a number of low level cues (e.g., color shifts, shot rhythm, music onset). In all of these efforts, results of image processing serve as important guides to select keyframes for browsing or summary presentations. Pentland's chapter concludes this third section by aiming at deeper semantic representations of video content of humans (e.g., tracking heads, hands, feet; classifying facial expressions).

3.4 Speech and Language Processing for Video Access

The fourth section of the book addresses content based access to video via indices of associated streams of spoken and written language (e.g. closed caption text). The first chapter by Jones et al. report techniques for video mail retrieval using spoken language indexing. Hauptman and Witbrock also investigate large vocabulary, continuous speaker independent broadcast news transcription, however for broadcast news retrieval and skimming. The last two chapters of this section turn to language processing of text transcripts. Mani et al. segment broadcast news using discourse cue based processing to enable story-based browsing of broadcast news. Takeshita et al. similarly report on discourse and lexical language processing to support "semantic based skim structures".

3.5 Architectures and Tools

Developing reusable architectures and tools is an important development in any field and is the focus of the fifth section. Whereas primary emphasis has been placed on algorithmic development, given the complexity of media analysis, researchers have begun to address the issues of developing frameworks and facilities to support both integration and analysis of multimedia data and processing. In the first chapter, Mérialdo and Dubois describe their Multimedia Flow Browser and related Agent Editor that enables users to both visualize agent interactions and combine existing agents into new agents. In the second chapter, Adams et al. present MOODS, a framework for developing content-based retrieval applications that allow users to go beyond searching for features such as colors and textures by combining a database, processing engine, and knowledge base. Their framework is illustrated in its application to music note recognition and ancient manuscript analysis. In addition to tools and frameworks to control and coordinate processing, multimedia systems also require multimedia query and analysis tools that go beyond the visual query reported in earlier sections. In the final chapter of this section, Hibino and Rundensteiner report their development of a visual temporal query mechanism and a visualization tool that supports temporal analysis of data from computer supported cooperative work environments.

3.6 Intelligent Hypermedia Access

Chapters in the sixth section of the book shift from indexing material to improving user access through human computer interfaces that support more effective multimedia interaction. The vision of conversational multimodal access necessarily entails user models and discourse models, and their application to either adapting hypertext or more directly managing interaction.

In the first chapter, Kobsa et al. describe work on adaptive hypertext and hypermedia systems that are tailored to a users' knowledge, interests and abilities, relying upon the services of a networked user modeling shell. In the second chapter, Vassileva presents a user and role adaptive, task-based hypermedia interface and shows how this improves the performance of both novice and expert users. The final two chapters both deal with multimedia dialogue in the context of art information access. In the first, Stock et al. describe the integration of a mediated information access paradigm (based on natural language dialog) and a navigational one (exploiting hypermedia), incorporating the use of attentional state and associated communicative acts to improve the effectiveness of multimodal information access. Finally, Stein et al. present a conversational approach to interactive retrieval which allows for an active role of the system, employing abductive reasoning to interpret ambiguous queries and to maintain coherent dialogue.

3.7 Empirical Studies

The final section of the book addresses empirical studies of multimedia information retrieval systems which aim at a deeper understanding of the strengths and weaknesses of various media and the way users interact with these. In the first chapter in this section, Horacek reports results of experiments with users of multimedia on-line documentation from which he derives design guidelines for object and action descriptions and information organization (e.g., “make cross media references explicit”). In the final chapter, Sutcliffe et al. report empirical studies of multimedia information retrieval which show that users may be misled into searching inappropriate media by the way a question is expressed, how explicit cross references between media can help in the extraction of information, and that well known information retrieval problems such as null result sets are exacerbated by multimedia. Like Horacek, the authors present design guidelines to overcome these challenges.

3.8 Content Index

Because many of the chapters in this collection address issues which cut across the above section distinctions, in order to facilitate access for the reader, Table 1 cross references chapters by:

1. Media data types investigated (e.g., text (captions or transcriptions), speech, sounds or music, graphics, imagery, video) and if the investigations examine cross stream or multiple media analysis.

2. The media application they address, e.g., indexing broadcast video news, video teleconferences, or video mail; supporting hypermedia information access.

3. Principle algorithmic/functionality issues addressed (e.g., multimedia indexing, browsing, searching).

4. Reported techniques, such as statistical-, rule-, case-, model-, or agent-based processing; the application of user or discourse models to tailor interaction.

For example, if a reader is interested only in techniques for multimedia processing exploiting text transcriptions, Figure 1 points to Chapters 5-8 , but also leads the reader to Chapter 13 which deals with integrated processing of text captioned images, as well as Chapters 11 and 15, which deal with image processing of text manuscripts. All chapters address, from an algorithmic or analytic perspective, multimedia information access (e.g., to support browsing, search, or extraction).

CHAPTER	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Text			v	v						v	v	v	v		v							
Speech										v	v	v		v								
Sound						v		v														
Graphics					v																	
Imagery	v	v	v	v					v						v							
Video	v						v	v	v	v	v	v	v	v		v						
Multistream								v			v	v	v	v								
News Access							v				v	v		v								
Teleconference																v						
Video Mail										v												
Documentation																	v				v	
Hypermedia																	v	v	v	v	v	v
Indexing	v	v	v	v	v	v	v	v	v	v	v	v	v	v	v							
Search	v	v	v	v	v	v	v	v	v	v	v	v	v	v	v	v	v	v	v	v	v	v
Browse	v	v		v	v		v	v				v	v	v								
Visualization																v						
Statistical	v	v	v	v		v	v		v	v	v											
Rule based								v														
Case based				v																		
Model based	v		v		v				v	v	v						v	v	v	v		
Agent based														v								
User Models																	v	v	v			
Discourse												v	v						v	v		

Table 1. Content Index of Chapters

4. Key Remaining Problems

Whereas this collection presents techniques for graphics, audio, imagery, and video information detection, extraction, and/or interaction, many problems must be solved in order to enable multimedia indexing, search, and navigation for large scale, heterogeneous collections. The most significant remaining systems level challenges include:

1. *Scalability and Performance*: Dealing with orders of magnitude larger volumes of multimedia information, that support (storage and time) efficient indexing and real-time retrieval.

2. *Portability*: Creating algorithms which rely minimally on domain-specific knowledge and can rapidly be applied to new multimedia corpora. Unfortunately, there remains a dearth of multimedia (e.g., audio, imagery, video) annotated data sets for experimentation. These data sets are fundamental to be used as training and test sets for machine learned indexing algorithms. And because of their creation expense, we require tools to support efficient corpora creation and annotation (either automated, semi-automated, or manual).

3. *Robustness*: Dealing with increasing levels of irrelevance, dirty data and unstructured data.

4. *Extensibility*: As new processing algorithms emerge (e.g., for indexing, search, extraction, summarization), system architectures should be sufficiently flexible to support seemless integration with existing approaches (e.g., augmenting a video indexing system with a speaker identification or face identification algorithm).

5. *Usability.* Increasingly sophisticated processing will drive a need for mechanisms that can mitigate complexity to ensure natural and learnable interfaces that ameliorate and do not exacerbate task accomplishment.

We next consider challenges in the active research areas of media representation, reasoning (including algorithms), interaction, and evaluation.

4.1 Representation

One fundamental issue is the definition of primitives for the representation of data, information, and knowledge within and across different media, such as video, speech and audio. Media dependent and independent indices must be created to enable seemless cross media integration in order to support multimedia browsing, search, extraction and summarization. Figure 3 above alludes to some common ground. However, the designer of such representations needs to be aware of several distinct, often conflicting, objectives. There will be tradeoffs between completeness and timeliness of processing, expressiveness and compactness of representations, and computability and communicability of indices (i.e., if necessary, is the representation intuitive to the end-user).

Unfortunately, in many cases the indices automatically generated by current methods (e.g., shape, color, texture for imagery) are not rich enough to support more advanced processing, such as media understanding and generation. For example, to support content based audio access, Blum et al. (this volume) found the need to represent both measurable acoustic/perceptual properties (e.g., brightness, pitch, loudness) as well as subjective, user supplied attributes (e.g., "a shimmering" or "buzzing" sound). In many cases we must integrate heterogeneous representations. For example, how do we reconcile statistical image processing which involves representations at the depth of perceptual properties (e.g., color, shape, size, motion vectors) with the semantic or object properties that language processing systems yield from principally symbolic processing (e.g., named entities such as people, places, and things, their relationships, and associated events)? Additional remaining research questions include, What are the elements of media and mode (e.g., languages for visual, auditory, tactile primitives), and what are their associated syntax, semantics, and pragmatics?, What are appropriate formal languages in which to capture these? Does a media interlingua exist?, How can efficient but effective indices be organized (e.g., as a hybrid of statistical and semantic elements, as a casebase)?, Will these representations be scalable, for example, to millions of cases?

4.2 Reasoning and Architectures

Assuming we understand how to represent media, what algorithms or methods will prove valuable for browsing, search, extraction, and summarization? Techniques for media segmentation, classification, and parsing include purely statistical ones as well as symbolic ones, using rules, cases, and models. While research has primarily focused on single channel analysis (e.g., the image, audio, or closed-caption streams of a video), researchers are now beginning to investigate cross-channel analysis. Also, as Fickner et al.'s chapter points out, there exist indexing techniques for exact matching or range searching of tabular data which assure sublinear search for indexing (e.g., binary trees).

However, what is required are fast and storage efficient techniques that support queries that include similarity matching together with spatial constraints. Finally, in part because of an inbalance between media analysis and generation techniques, there is a general lack of reversable methods, that is, algorithms that support both media analysis and generation. To be a good reader you must be a good writer and vice versa.

In addition to challenges with processing individual media, multiple media reveal special challenges such as the need for multistream alignment of heterogeneous data. However, these can also be viewed as opportunities, in which exploitation of cross channel cues can yield enhanced algorithmic performance as shown in several chapters in this volume (e.g., Aigraine et al.'s multiple cue macrostructure detector, Rowe et al.'s use of captions to improve image classification, Mani et al.'s demonstration of the value of speaker ID to improve text-based topic segmentation). Griffioen et al.'s framework approach and Mérialdo and Dubois' agent based approaches are mechanisms that can enable this kind of integrated processing.

Strongly related to these processes is the issue of what are the basic building blocks and processes within multimedia information retrieval systems (e.g., for indexing, searching, merging media elements). From an architectural standpoint, we need to understand a number of issues, including: What are the key components?, What functionality do they need to support?, What processes are most appropriate for each of these functions (e.g., statistically-based, knowledge-based)?, What are the appropriate integrating architectures (e.g., distributed vs. centralized, use of agent technology)?, What is the proper flow of control?, and How should they interact (e.g., serially, interleaved, co-constraining)?, What is the appropriate degree of automation vs. manual annotation/intervention?, How general will these architectures be across multiple applications and domains?

4.3 Interaction with Multimedia Information

Interaction tailored to the user, task, and situation will be necessary to support interaction with large scale and complex digital libraries. Heterogeneous media sources/services may require different methods of access, including distinct query languages and profiles (e.g., keywords for text data, structured query for relational data, visual query), however these should be as intuitive and uniform as possible, supporting cross media query. The notion of integration is important but this can mean different things: at one end of the spectrum it can mean a single, integrated representation system working across different media (a challenging and perhaps impossible objective, as discussed in Section 4.1 above); at the other end it can mean an integrated modus vivendi for the user at her terminal to enable shifting easily from one medium to another.

Important multimedia information interaction research questions include: What is the role of the user model in a multimedia interactive context? What is the role of discourse structure and multimedia structure? What is the role of managing models of attention, including global and local context? Is there a role for tasks if the domain or information is not defined or ill-defined? Is information negotiation a useful concept in an exploration environment? What common communicative devices found in human communication appear in multimedia (e.g., discourse context, anaphora, co-reference, elision)? What will be the requirements and supporting tools and techniques for automated presentation design? What are effective multimedia displays and/or visualization/browsing of time-based media and in differing information seeking tasks? How can automated agents support and mediate conversational interaction to move toward a digital information retrieval assistant?

4.4 Evaluation

Finally, we need to better understand from an empirical standpoint how well our multimedia information retrieval systems will function. This entails designing metrics and conducting evaluations, often contrasting human and machine performance or integrated human-machine performance. While traditional metrics of precision and recall will remain important in information seeking contexts, we need to develop evaluation approaches which include humans in the loop.

Unfortunately, there remains a limited amount of reported experimentation in this area. Whereas scientific endeavors in areas such as vision, speech recognition, and text processing have made progress by creating large scale training and test corpora (e.g., the TIPSTER text retrieval and message understanding collections), collection and annotation of multimedia corpora is expensive and remains a bottleneck. Unfortunately, without such corpora, many issues of scalability, portability, extensibility and robustness cannot be properly investigated. As a consequence of the lack of large scale multimedia corpora, much of the current experimentation results in important but merely indicative conclusions, as discussed by several authors in this collection, including Jones et al. and Mani et al.

In addition to multimedia data, we need carefully defined multimedia tasks and experiments. Researchers have suggested the need for explicit task definitions (i.e., well-defined user and context centered goals), explicit separation of media, and the application of the scientific method in which one adds or subtracts dependencies (e.g., media, tasks, user classes) and runs controlled experiments where the goal is to lead to usability and performance criteria by comparison with some baseline (e.g., versus use of a conventional library). It is also necessary to have not only well defined but also large scale evaluations (e.g. hundreds of queries and relevancy judgments, gigabytes of data) with the ultimate objective of discerning principles of multimedia interaction which can be both prescriptive and predictive.

4.5 General Issues

As with any new technology, there are a range of economic and social challenges beyond the above technical ones that will arise in the context of multimedia information retrieval. On one hand, the technology promises information retrieval that can be more natural and personalized while at the same time more exact and richer in content and form. On the other, the power of information manipulation raises legal and economic issues (e.g., how to track and price derived works such as an extraction or summary), social issues (e.g., how to ensure all citizens have equal access to multimedia data and advanced tools), and privacy issues (e.g., how to manage models of a user and their information retrieval interactions), although these are not necessarily unique to this research area. Traditional concepts such as intellectual property and copyright will require new operational definitions as slight modifications of content or form using media editing tools might dramatically alter the message of a presentation and, by some, be considered novel artifacts.

5. Conclusion

Many open research issues remain in this nascent, interdisciplinary area. Lessons learned from more long standing research communities (e.g., information retrieval, image processing, speech processing, language processing) regarding user-centered design, corpora development, and evaluation strategies can be leveraged to make more rapid progress in research and applications. Only by addressing the above key questions will underlying progress be made to support future applications such as video archive search, video and audio teleconferencing archiving, and content-based multimedia access. A key challenge will be the development and transition of techniques to the many domains with needs for multimedia archive access including medical records, retail marketing, stock photo and video management (e.g., for advertising), museum and library collections, scientific applications (e.g., environmental analysis, weather prediction), law enforcement (e.g., face, voice, handwriting recognition), and content based clip art (e.g., images, sounds, graphics).

If successful, intelligent multimedia information retrieval will improve information access in three principal ways. First, it promises more effective access to content: getting the right stuff and tailoring it to the context of the user, their task, and their environment. The goal of achieving context sensitivity will be limited only by the richness of models that can be created. Second, by providing only the most relevant information, together with high performance browsing and search tools, this not only enables more comprehensive and higher quality search, it saves time, thus saving money. Finally, enabling the user to ask for and receive information in a natural manner (e.g., by speaking, drawing, or pointing to similar artifacts) promises a less stressful and more pleasant interaction.

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 for an earlier concept for Figure 2, Wolfgang Wahlster, Philippe Aigraine, and Stephen Smoliar.

7. References

Baecker, R. M.; Grudin, J.; Buxton, W.; and Greenberg, S., eds. 1995. second edition. Readings in Human-Computer Interaction: Toward the Year 2000. San Mateo, CA: Morgan Kaufmann.

Blattner, M. and Dannenberg, R., eds. 1992. *Multimedia Interface Design*, ACM Press.

Chen, C. H.; Pau, L. F.; and Wang, P. S. P., eds. 1993. *Handbook of Pattern Recognition and Computer Vision.* Singapore: World Scientific.

Furht, B.; Smoliar, S.; and Zhang, H. J., eds. 1996. *Video and Image Processing in Multimedia Systems*. Boston: Kluwer.

Furht, B. ed. 1996. *Multimedia Systems and Techniques*. Boston: Kluwer.

Grosz, B. J., Sparck Jones, K., and Webber, B., eds. 1986. *Readings in Natural Language Processing.* Los Altos, CA: Morgan Kaufmann.

Grosz, B. J. and C. Sidner. July-September, 1986. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics* 12(3): 175-204.

IFIP, 1989 and 1992. *Visual Database Systems I and II*, North-Holland: Elsevier Science Publishers.

Maybury, M. T. 1991. Topical, Temporal and Spatial Constraints on Linguistic Realization. In Pattabhiraman, T. and Cercone, N., eds. *Computational Intelligence: Special Issue on Natural Language Generation.* 7(4): 266-275.

Maybury, M. T. 1993. *Intelligent Multimedia Interfaces*. Menlo Park, CA/Cambridge, MA: AAAI/MIT Press.

MUC-6, Proceedings of the Sixth Message Understanding Conference. Advanced Research Projects Agency Information Technology Office, Columbia, MD, 6-8 November, 1995.

Niblack, W. and Jain, R., eds. 1993, 1994, 1995. *Proceedings of IS&T/SPIE. Conference on Storage and Retrieval for Image and Video Databases I, II, and III*, Vols. 1908, 2185, and 2420. Bellingham, WA: SPIE.

van Rijjsbergen, C. J. 1979. *Information Retrieval*. London: Butterworths.

Salton, G. *Automatic Text Processing*. Reading, MA: Addison-Wesley, 1988.

Sullivan, J. and Tyler, S., eds. 1991. *Intelligent User Interfaces*. Reading, MA: Addison-Wesley, ACM Press.

Waibel, A. and Lee, K., eds. 1990. *Readings in Speech Recognition*, San Mateo, CA: Morgan Kaufmann.