CHAPTER 1

INTRODUCTION

MARK T. MAYBURY

1.1 MOTIVATION

Our world has become massively multimedia. In addition to rapidly growing personal and industrial collections of music, photography, and video, media sharing sites have exploded in recent years. The growth of social media sites for not only social networking but for information sharing has further fueled the broad and deep availability of media sources. Even special industrial collections once limited to proprietary access (e.g., Time-Life images), or precious books or esoteric scientific materials once restricted to special collection access, or massive scientific collections (e.g., genetics, astronomy, and medical), or sensors (traffic, meteorology, and space imaging) once accessible only to a few privileged users are increasingly becoming widely accessible.

Rapid growth of global and mobile telecommunications and the Web have accelerated both the growth of and access to media. As of 2012, over one-third of the world's population is currently online (2.3 billion users), although some regions of the world (e.g., Africa) have less than 15% of their potential users online. The World Wide Web runs over the Internet and provides easy hyperlinked access to pages of text, images, and video—in fact, to over 800 million websites, a majority of which are commercial (.com). The most visited site in the world, Google (Yahoo! is second) performs hundreds of millions of Internet searches on millions of servers that process many petabytes of user-generated content daily. Google has discovered over one trillion unique URLs. Wikis, blogs, Twitter, and other social media (e.g., MySpace and LinkedIn) have grown exponentially. Professional imagery sharing on Flickr now contains over 6 billion images. Considering social networking, more than 6 billion photos and more than 12 million videos are uploaded each

Multimedia Information Extraction: Advances in Video, Audio, and Imagery Analysis for Search, Data Mining, Surveillance, and Authoring, First Edition. Edited by Mark T. Maybury.

month on Facebook by over 800 billion users. Considering audio, IP telephony, pod/broadcasting, and digital music has similarly exploded. For example, over 16 billion songs and over 25 billion apps have been downloaded from iTunes alone since its 2003 launch, with as many as 20 million songs being downloaded in one day. In a simple form of extraction, loudness and frequency spectrum analysis are used to generate music visualizations.

Parallel to the Internet, the amount of television consumption in developed countries is impressive. According to the A.C. Nielsen Co., the average American watches more than 4 hours of TV each day. This corresponds to 28 hours each week, or 2 months of nonstop TV watching per year. In an average 65-year lifespan, a person will have spent 9 years watching television. Online video access has rocketed in recent times. In April of 2009, over 150 million U.S. viewers watched an average of 111 videos watching on average about six and a half hours of video. Nearly 17 billion online videos were viewed in June 2009, with 40 percent of these at Youtube (107 million viewers, averaging 3–5 minutes each video), a site at which approximately 20 hours of video are uploaded every minute, twice the rate of the previous year. By March 2012, this had grown to 48 hours of video being uploaded every minute, with over 3 billion views per day. Network traffic involving YouTube accounts for 20% of web traffic and 10% of all Internet traffic. With billions of mobile device subscriptions and with mobiles outnumbering PCs five to one, increasingly access will be mobile. Furthermore, in the United States, four billion hours of surveillance video is recorded every week. Even if one person were able to monitor 10 cameras simultaneously for 40 hours a week, monitoring all the footage would require 10 million surveillance staff, roughly about 3.3% of the U.S. population. As collections of personal media, web media, cultural heritage content, multimedia news, meetings, and others develop from gigabyte to terabyte to petabyte, the need will only increase for accurate, rapid, and cross-media extraction for a variety of user retrieval and reuse needs. This massive volume of media is driving a need for more automated processing to support a range of educational, entertainment, medical, industrial, law enforcement, defense, historical, environmental, economic, political, and social needs.

But how can we all benefit from these treasures? When we have specific interests or purposes, can we leverage this tsunami of multimedia to our own individual aims and for the greater good of all? Are there potential synergies among latent information in media awaiting to be extracted, like hidden treasures in a lost cave? Can we infer what someone was feeling when their image was captured? How can we automate currently manually intensive, inconsistent, and errorful access to media? How close are we to the dream of automated media extraction and what path will take us there?

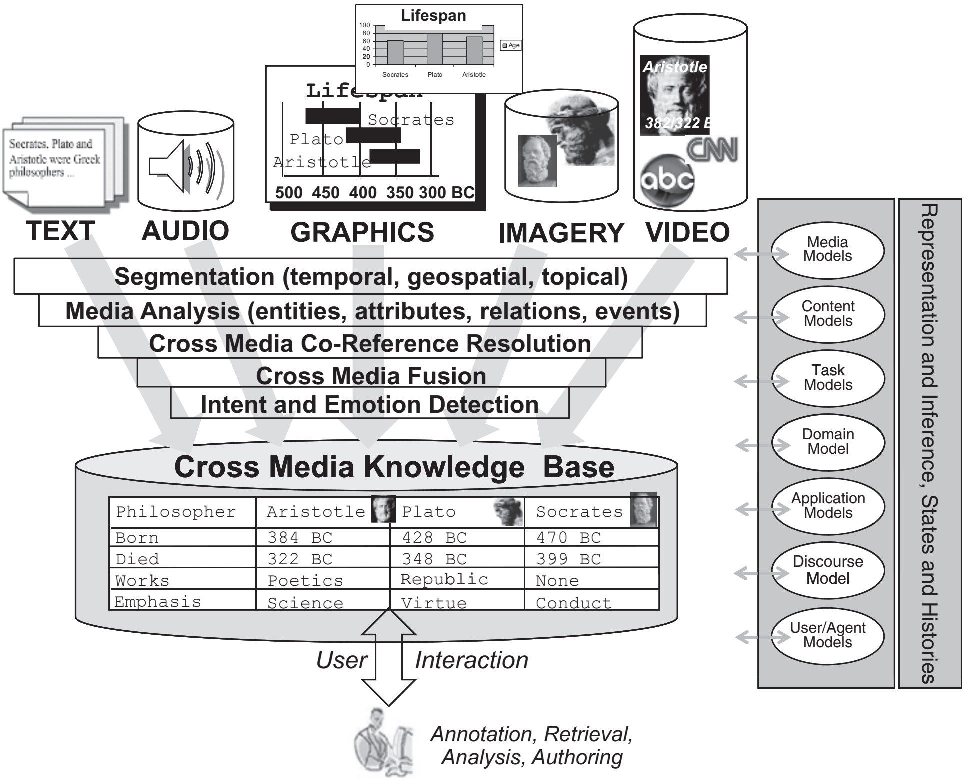
This collection opens windows into some of the exciting possibilities enabled by extracting information, knowledge, and emotions from text, images, graphics, audio, and video. Already, software can perform content-based indexing of your personal collections of digital images and videos and also provide you with content-based access to audio and graphics collections. And analysis of print and television advertising can help identify in which contexts (locations, objects, and people) a product appears and people's sentiments about it. Radiologists and oncologists are beginning to automatically retrieve cases of patients who exhibit visually similar conditions in internal organs to improve diagnoses and treatment. Someday soon, you will be able to film your vacation and have not only automated identification of the people and places in your movies, but also the creation of a virtual world of reconstructed people, objects, and buildings, including representation of the happy, sad, frustrating, or exhilarating moments of the characters captured therein. Indeed, multimedia information extraction technologies promise new possibilities for personal histories, urban planning, and cultural heritage. They might also help us better understand animal behavior, biological processes, and the environment. These technologies could someday provide new insights in human psychology, sociology, and perhaps even governance.

The remainder of this introductory chapter first defines terminology and the overall process of multimedia information extraction. To facilitate the use of this collection in research, it then describes the collection's structure, which mirrors the key media extraction areas. This is augmented with a hierarchical index at the back of the book to facilitate retrieval of key detailed topics. To facilitate the collection's use in teaching, this chapter concludes by illustrating how each section addresses standard computing curricula.

1.2 DEFINITIONS

Multimedia information extraction is the process of analyzing multiple media (e.g., text, audio, graphics, imagery, and video) to excerpt content (e.g., people, places, things, events, intentions, and emotions) for some particular purpose (e.g., data basing, question answering, summarization, authoring, and visualization). Extraction is the process of pulling out or excising elements from the original media source, whereas abstraction is the generalization or integration across a range of these excised elements (Mani and Maybury 1999). This book is focused on the former, where extracted elements can stand alone (e.g., populating a database) or be linked to or presented in the context of the original source (e.g., highlighted named entities in text or circled faces in images or tracked objects moving in a video).

As illustrated in Figure 1.1, multimedia information extraction requires a cascading set of processing, including the segmentation of heterogeneous media (in terms of time, space, or topic), the analysis of media to identify entities, their properties and relations as well as events, the resolution of references both within and across media, and the recognition of intent and emotion. As is illustrated on the right hand side of the figure, the process is knowledge intensive. It requires models of each of the media, including their elements, such as words, phones, visemes, but also their properties, how these are sequenced and structured, and their meaning. It also requires the context in which the media occurs, such as the time (absolute or relative), location (physical or virtual), medium (e.g., newspaper, radio, television, and Internet), or topic. The task being performed is also important (its objective, steps, constraints, and enabling conditions), as well as the domain in which it occurs (e.g., medicine, manufacturing, and environment) and the application for which it is constructed (e.g., training, design, and entertainment). Of course, if the media extraction occurs in the context of an interaction with a user, it is quite possible that the ongoing dialogue will be important to model (e.g., the user's requests and any reaction they provide to interim results), as well as a model of the user's goals, objectives, skills, preferences, and so on. As the large vertical arrows in the figure show, the processing of each media may require unique algorithms. In cases where multiple


Lifespan
100
80
60
40
20
0
Age
Socrates
Plato
Aristotle
Lifespan
Socrates
Plato
Aristotle
500 450 400 350 300 BC
Socrates, Plato and Aristotle were Greek philosophers ...
Aristotle
382/322 B
CNN
abc
TEXT
AUDIO
GRAPHICS
IMAGERY
VIDEO
Segmentation (temporal, geospatial, topical)
Media Analysis (entities, attributes, relations, events)
Cross Media Co-Reference Resolution
Cross Media Fusion
Intent and Emotion Detection
Cross Media Knowledge Base
User
Interaction
Annotation, Retrieval, Analysis, Authoring
Media Models
Content Models
Task Models
Domain Model
Application Models
Discourse Model
User/Agent Models
Representation and Inference, States and Histories

Philosopher	Aristotle	Plato	Socrates
Born	384 BC	428 BC	470 BC
Died	322 BC	348 BC	399 BC
Works	Poetics	Republic	None
Emphasis	Science	Virtue	Conduct

Figure 1.1. Multimedia information extraction.

media contain synchronous channels (e.g., the audio, imagery, and on screen text in video broadcasts), media processing can often take advantage of complementary information in parallel channels. Finally, extraction results can be captured in a cross-media knowledge base. This processing is all in support of some primary user task that can range from annotation, to retrieval, to analysis, to authoring or some combination of these.

Multimedia information extraction is by nature interdisciplinary. It lies at the intersection of and requires collaboration among multiple disciplines, including artificial intelligence, human computer interaction, databases, information retrieval, media, and social media studies. It relies upon many component technologies, including but not limited to natural language processing (including speech and text), image processing, video processing, non-speech audio analysis, information retrieval, information summarization, knowledge representation and reasoning, and social media information processing. Multimedia information extraction promises advances across a spectrum of application areas, including but not limited to web search, photography and movie editing, music understanding and synthesis, education, health care, communications and networking, and medical sensor exploitation (e.g., sonograms and imaging).

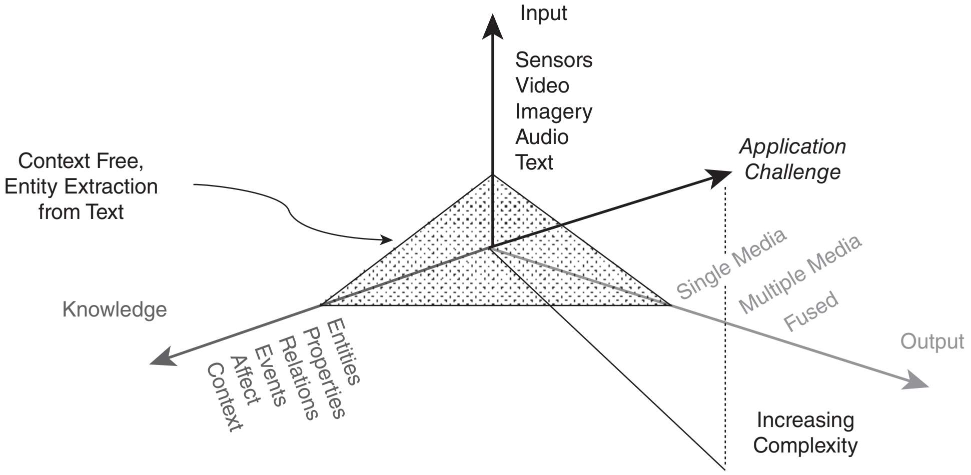


Figure 1.2. Some dimensions of multimedia information extraction.

As Figure 1.2 illustrates, multimedia information extraction can be characterized along several dimensions, including the nature of the input, output, and knowledge processed. In terms of input, the source can be single media, such as text, audio, or imagery; composite media, such as video (which includes text, audio, and moving imagery); wearable sensors, such as data gloves or bodysuits, or remote sensors, such as infrared or multispectral imagers; or combinations of these, which can result in diverse and large-scale collections. The output can range from simple annotations on or extractions from single media and multiple media, or it can be fused or integrated across a range of media. Finally, the knowledge that is represented and reasoned about can include entities (e.g., people, places, and things), their properties (e.g., physical and conceptual), their relationships with one another (geospatial, temporal, and organizational), their activities or events, the emotional affect exhibited or produced by the media and its elements, and the context (time, space, topic, social, and political) in which it appears. It can even extend to knowledge-based models of and processing that is sensitive to the domain, task, application, and user. The next chapter explores the state of the art of extraction of a range of knowledge from a variety of media input for various output purposes.

Figure 1.3 steps back to illustrate the broader processing environment in which multimedia information extraction occurs. While the primary methods reported in this collection address extraction of content from various media, often those media will contain metadata about their author, origin, pedigree, contents, and so on, which can be used to more effectively process them. Similarly, relating one media to another (e.g., a written transcript of speech, an image which is a subimage of another image) can be exploited to improve processing. Also, external semi-structured or structured sources of data, information, or knowledge (e.g., a dictionary of words, an encyclopedia, a graphics library, or ontology) can enhance processing as illustrated in Figure 1.3. Finally, information about the user (their knowledge, interests, or skills) or the context of the task can also enhance the kind of information that is extracted or even the way in which it is extracted (e.g., incrementally or in batch

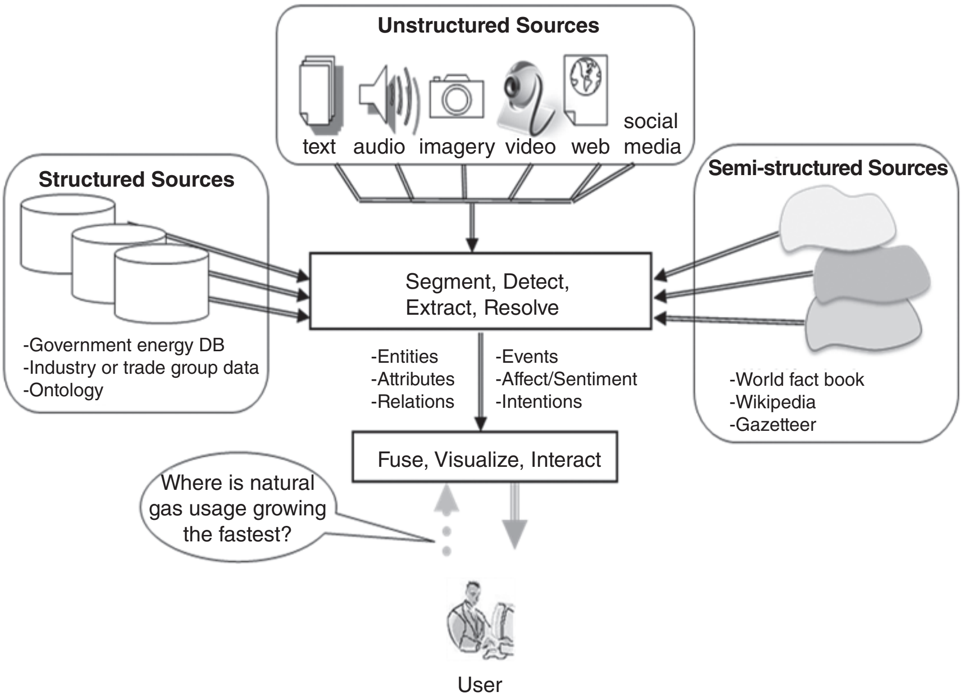


Figure 1.3. Multimedia architecture.

mode). Notably, the user's question itself can be multimedia and may require multimedia information extraction during query processing.

1.3 COLLECTION OVERVIEW

The five sections of *Multimedia Information Extraction* represent key areas of research and development including audio, graphics, imagery, and video extraction, affect and behavior extraction, and multimedia annotation and authoring.

1.3.1 Section 1: Image Extraction

Exponential growth of personal, professional, and public collections of imagery requires improved methods for content-based and collaborative retrieval of whole and parts of images. This first section considers the extraction of a range of elements from imagery, such as objects, logos, visual concepts, shape, and emotional faces. Solutions reported in this section enable improved image collection organization and retrieval, geolocation based on image features, extraction of 3D models from city or historic buildings, and improved facial emotion extraction and synthesis. The chapters identify a number of research gap areas, including image query context, results presentation, and representation, and reasoning about visual content.

1.3.2 Section 2: Video Extraction

The rapid growth of digital video services and massive video repositories such as YouTube provide challenges for extraction of content from a broad range of video domains from broadcast news to sports to surveillance video. Solutions reported in this section include how processing of the text and/or audio streams of video can improve the precision and recall of video extraction or retrieval. Other work automatically identifies bias in TV news video through analysis of written words, spoken words, and visual concepts that reflect both topics and inner attitudes and opinions toward an issue. Tagging video with multiple viewpoints promises to foster better informed decisions. In other applied research, global access to multilingual video news requires integration of a broad set of image processing (e.g., keyframe detection, face identification, scene cut analysis, color frame detection, on screen OCR, and logo detection), as well as audio analysis (e.g., audio classification, speaker identification, automatic speech recognition, named entity detection, closed captioning processing, and machine translation). Performance can be enhanced using cross media extraction, for example, correlating identity information across face identification, speaker identification, and visual OCR. In the context of football game processing, another chapter considers speech and language processing to detect touchdowns, fumbles, and interceptions in video. The authors are able to detect banners and logos in football and baseball with over 95% accuracy. Other solutions provide detection and recognition of text content in video (including overlaid and in-scene text). Notably, a majority of entities in video text did not occur in speech transcripts, especially location and person names and organization names. Other solutions do not look at the content but rather frequency of use of different scenes in a video to detect their importance. Yet a different solution considers anomaly detection from uncalibrated camera networks for tasks such as surveillance of cars or people. Overall, the chapters identify a number of research gap areas, such as the need for inexpensive annotation, cross-modal indicators, scalability, portability, and robustness.

1.3.3 Section 3: Audio, Graphics, and Behavior Extraction

Media extraction is not limited to traditional areas of text, speech, or video, but includes extracting information from non-speech audio (e.g., emotion and music), graphics, and human behavior. Solutions reported in this section include identity, content, and emotional feature audio extraction from massive, multimedia, multilingual audio sources in the audio hot spotting system (AHS). Another chapter reports extraction of information graphics (simple bar charts, grouped bar charts, and simple line graphs) using both visual and linguistic evidence. Leveraging eye tracking experiments to guide perceptual/cognitive modeling, a Bayesian-based message extractor achieves an 80% recognition rate on 110 simple bar charts. The last chapter of the section reveals how "thin slices" of extracted social behavior fusing nonverbal cues, including prosodic features, facial expressions, body postures, and gestures, can yield reliable classification of personality traits and social roles. For example, extracting the personality feature "locus of control" was on average 87% accurate, and detecting "extraversion" was on average 89% accurate. This section reveals important new frontiers of extracting identity and emotions, trends and relationships, and personality and social roles.

1.3.4 Section 4: Affect Extraction from Audio and Imagery

This section focuses on the extraction of emotional indicators from audio and imagery. Solutions described include the detection of emotional state, age, and gender in TV and radio broadcasts. For example, considering hundreds of acoustic features, whereas speaker gender can be classified with more than 90% accuracy, age recognition remains difficult. The correlation coefficient (CC) between the best algorithm and human (where 1 is perfect correlation) was 0.62 for valence (positive vs. negative) and 0.85 for arousal (calm vs. excited) traits. Another chapter explores valenced (positive/negative) expressions, as well as nonlinguistic reactions (e.g., applause, booing, and laughter) to discover their importance to persuasive communication. Valenced expressions are used to distinguish, for example, Democrat from Republican texts with about 80% accuracy. In contrast, considering images annotated with induced emotional state in the context of systems such as Flickr and Facebook, 68% of users are provided better (in terms of precision) recommendations through the use of affective metadata. The last chapter of the section reports the extraction of low-level, affective features from both the acoustic (e.g., pitch, energy) and visual (e.g., motion, shot cut rate, saturation, and brightness) streams of feature films to model valence and arousal. Scenes with particular properties are mapped to emotional categories, for example, a high-pitched human shouting with dark scenes might indicate horror or terror scenes, whereas those with bright colors might indicate funny or happy scenes. Together, these chapters articulate the emergence of a diversity of methods to detect affective features of emotion in many media, such as audio, imagery, and video.

1.3.5 Section 5: Multimedia Annotation and Authoring

This final section turns to methods and systems for media annotation and authoring. Solutions include the more precise annotation of human movement by extending the publicly available ANVIL (http://www.anvil-software.de) to perform 3D motion capture data annotation, query, and analysis. Another chapter employs a display grammar to author and manage interactions with imagery and extracted audio. A related chapter demonstrates how semantic query-based authoring can be used to design interactive narratives, including 2D images, sounds, and virtual camera movements in a 3D environment about historical Brooklyn. Ontology-based authoring supports concept navigation among an (SoundFisher system) audio analysis of non-speech natural sounds. The last chapter of the section describes the MADCOW system for annotation of relations in multimedia web documents. One unique feature of MADCOW is the ability to add annotations not only to single but also multiple portions of a document, potentially revealing new relations. By moving media assembly to the point of delivery, users' preferences, interests, and actions can influence display.

1.4 CONTENT INDEX

This content index is intended for researchers and instructors who intend to use this collection for research and teaching. In order to facilitate access to relevant content, each chapter is classified in Table 1.1 according to the type of media it addresses,

TABLE 1.1. Content Index of Chapters

Section	Chapter	Media Type					Task Application							Architecture Emphasis				Technical Approach			
		Text	Audio	Imagery	Graphics	Video	Web Access/Exploit	Image Processing/Mgmt	Facial Processing/Mgmt	Broadcast News Video	Meetings	Surveillance	Affect Detection	Annotation	Extraction	Retrieval	Authoring	Statistical	Symbolic/Model Based	Recommender	Social
	2	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■				
I	3			■	■			■							■	■		■			
	4			■			■	■								■		■			
	5			■	■			■							■			■	■		
	6			■		■			■				■		■			■			
II	7			■				■							■			■			■
	8					■				■		■			■	■		■			
	9	■	■			■				■		■	■		■	■		■			
	10	■	■			■				■		■			■	■		■			
	11					■				■					■	■		■			
	12					■	■								■	■		■			■
	13			■		■						■			■	■		■	■		
III	14		■			■				■	■	■			■	■		■	■		
	15				■		■	■							■			■			
	16		■	■		■					■		■		■			■			
IV	17	■	■			■				■			■		■	■		■			
	18	■	■							■	■		■		■			■			■
	19			■				■					■			■		■		■	
	20		■			■				■	■		■			■		■			
V	21	■				■	■							■					■		
	22		■	■	■	■	■							■			■		■		
	23		■	■	■	■	■							■			■		■		
	24	■		■			■							■					■		

the application task addressed, its architectural focus, and the technical approach pursued (each shown in four main columns in the Table). The table first distinguishes the media addressed by each chapter, such as if the chapter addresses extraction of text, audio, imagery (e.g., Flickr), graphics (e.g., charts), or video (e.g., YouTube). Next, we characterize each chapter in terms of the application task it aims to support, such as World Wide Web access, image management, face recognition, access to video (from broadcast news, meetings, or surveillance cameras), and/or detection of emotion or affect. Next, Table 1.1 indicates the primary architectural focus of each chapter, specifying whether the research primarily explores media annotation, extraction, retrieval, or authoring. Finally, Table 1.1 classifies each chapter in terms of the technical approach explored, such as the use of statistical or machine learning methods, symbolic or model-based methods (e.g., using knowledge sources such as electronic dictionaries as exemplified by WordNet, ontologies, and inference machinery such as that found in CYC, and/or semi-structured information sources, such as Wikipedia or the CIA fact book), recommender technology, and, finally, social technology.

1.5 MAPPING TO CORE CURRICULUM

Having seen how the chapters relate to core aspects of multimedia, we conclude by relating the sections of this collection to the required body of knowledge for core curriculum in human computer interaction, computer science, and information technology. This mapping is intended to assist instructors who plan to use this text in their classroom as a basis for or supplement to an undergraduate or graduate course in multimedia. The three columns in Table 1.2 relate each section of the book (in rows) to the ACM SIGCHI HCI Curricula (SIGCHI 1996), the ACM/IEEE computer science curricula (CS 2008), and the ACM/IEEE information technology curricula (IT 2008). In each cell in the matrix, core topics are listed and electives are italicized. For example, the Association for Computing Machinery (ACM) and the IEEE Computer Society have developed a model curricula for computer science that contains core knowledge areas, such as Discrete Structures (DS), Human–Computer Interaction (HC), Programming Fundamentals (PF), Graphics and Visual Computing (GV), Intelligent Systems (IS), and so on. Some of these topics are addressed throughout the collection (e.g., human computer interaction), whereas others are specific to particular sections (e.g., geometric modeling). Finally, the NSF Digital Libraries Curriculum project (DL 2009) is developing core modules that many of the sections and chapters in the collection relate directly to, such as digital objects, collection development, information/knowledge organization (including metadata), user behavior/interactions, indexing and searching, personalization, and evaluation.

Moreover, there are many additional connections to core curricula that the individual instructor can discover based on lesson plans that are not captured in the table. For example, face and iris recognition, addressed in Chapter 2 and Chapter 6, is a key element in the ACM/IEEE core requirement of Information Assurance and Security (IAS). There are also Social and Professional Issues (SP) in the privacy aspects of multimedia information extraction and embedded in the behavior and affect extraction processing addressed in Sections 3 and 4. Finally, there are, of

TABLE 1.2. Book Sections Related to Core Curricula in HCI, CS, and IT (Electives Italicized)

Book Section	ACM SIGCHI Core Content	ACM/IEEE CS Core Curricula	ACM/IEEE IT Core Curricula
All Sections	User Interfaces, Communication and Interaction, Dialogue, Ergonomics, Human–Machine Fit and Adaptation	Human–Computer Interaction (HC), Discrete Structures (DS), Programming Fundamentals (PF), Algorithms and Complexity (AL) Intelligent Systems (IS), Information Management (IM), Net-Centric Computing (NC) Software Engineering (SE), Programming Languages (PL), *Multimedia Technologies, Machine Learning, Data Mining, Privacy and Civil Liberties*	Human Computer Interaction (HCI), Information Management (IM), Integrative Programming and Technologies (IPT), Math and Statistics for IT (MS), Programming Fundamentals (PF), *History of IT, Privacy and Civil Liberties, Digital Media*
Introduction and State of the Art	Evaluation	*Information Storage and Retrieval*	
I. Image Extraction	Image Processing	Graphics and Visual Computing (GV), *Perception, Geometric Modeling*	
II. Video Extraction	Audio/Image/Video Processing	Graphics and Visual Computing (GV), *Computer Vision*, *Natural Language Processing*	*Social Software*
III. Audio/Graphics/Behavior Extraction	Computer Graphics, Audio/Image/Video Processing	Graphics and Visual Computing (GV), *Natural Language Processing*	
IV. Affect Extraction in Audio and Video	Communication and Interaction, Audio/Image/Video Processing	*Signal Processing, Computer Vision*, *Natural Language Processing*	
V. Multimedia Annotation and Authoring	Input/Output Devices, Dialogue Techniques, Design	*Hypermedia, Multimedia and Multimodal Systems*	Web Systems and Technologies (WS)

course, System Integration & Architecture (SIA) challenges in creating a multimedia information extraction system that integrates text, audio, and motion imagery subsystems.

1.6 SUMMARY

This chapter introduces and defines multimedia information extraction, provides an overview of this collection, and provides both a content index and mapping of content to core curricula to facilitate research and teaching.

ACKNOWLEDGMENTS

Appreciation goes to all the authors for their feedback, especially to Björn Schuller and Gareth Jones for a careful reading of the final text.