Next-Generation Content Representation, Creation, and Searching for New-Media Applications in Education

SHIH-FU CHANG, MEMBER, IEEE, ALEXANDROS ELEFTHERIADIS, MEMBER, IEEE, AND ROBERT MCCLINTOCK, ASSOCIATE MEMBER, IEEE

Invited Paper

Content creation, editing, and searching are extremely time-consuming tasks that often require substantial training and experience, especially when high-quality audio and video are involved. "New media" represents a new paradigm for multimedia information representation and processing, in which the emphasis is placed on the actual content. It thus brings the tasks of content creation and searching much closer to actual users and enables them to be active producers of audio-visual information rather than passive recipients. We discuss the state of the art and present next-generation techniques for content representation, searching, creation, and editing. We discuss our experiences in developing a Web-based distributed compressed video editing and searching system (WebClip), a media-representation language (Flavor) and an object-based video-authoring system (Zest) based on it, and a large image/video search engine for the World Wide Web (WebSEEk). We also present a case study of new media applications based on specific planned multimedia education experiments with the above systems in several K–12 schools in Manhattan, NY.

Keywords*—*Content-based image/video search, content creation, content representation, multimedia education, new media application, video editing.

I. INTRODUCTION

Imagine how difficult intelligent writing would be if people always had to think about the form and shape of letters. Writing would be like chiseling inscriptions in stone, laborious and inflexible. Ideas and meaning would be opaque as attention fixed on the shape—"a curve, back, up, around, down, forward to the vertical plane where it started, straight up to the starting point, and then straight down to the horizontal plane of the lowest point on the curve." Such manipulations get one only a paltry indefinite article; how much more laborious inscribing any substantive noun or the action of a verb would be.

Work with images is still stranded in an analogous, primitive state where actions affect not visions, ideas, and thoughts but pixels on the screen. Our manipulation tools are so rudimentary that it is hard to think *with* an image, for to do anything we must think *about* the image. Given this state of the art, the potential value of digital media is still far from fulfilled. In education in particular, the student needs to grasp and master the content in question. Digital stonecutters make visualization programs to help ordinary people better understand complex ideas, but far too often, the complexity of digital tools for working with images makes them a distraction for the ordinary person seeking to express his thought.

A variety of intellectual functions are crucial to education and culture. Students need to learn how to store, retrieve, and cite materials of intellectual interest. They must create, edit, and manipulate challenging content. They must communicate, both receive and transmit, with others in an effort to sift and disseminate important ideas. All these functions are relatively well developed with the written resources of our culture. Vision is of immense intellectual power, but our imaging tools—tools to store, retrieve, and cite things we have seen; to create, edit, and manipulate meaningful images; and to receive, transmit, and disseminate them—are still far from fully developed.

Thus, educators need "new media" technology and applications, i.e., systems that go beyond mere digitization

Manuscript received August 2, 1997; revised December 8, 1997. The Guest Editor coordinating the review of this paper and approving it for publication was T. Chen. This work was supported by the AT&T Foundation and the industrial sponsors of Columbia University's ADVENT project. The work of S.-F. Chang was supported in part by the National Science Foundation under CAREER award (IRI-9 501 266) and STIMULATE award (IRI-9 619 124). The work of A. Eleftheriadis was supported in part by the National Science Foundation under CAREER award (MIP-9 703 163).

S.-F. Chang and A. Eleftheriadis are with the Department of Electrical Engineering, School of Engineering and Applied Science, Columbia New Media Technology Center, Columbia University, New York, NY 10027 USA (e-mail: sfchang@ee.columbia.edu; eleft@ee.columbia.edu).

R. McClintock is with the Institute for Learning Technologies, Teachers College, Columbia New Media Technology Center, Columbia University, New York, NY 10027 USA (e-mail: rom2@columbia.edu).

Publisher Item Identifier S 0018-9219(98)03379-9.

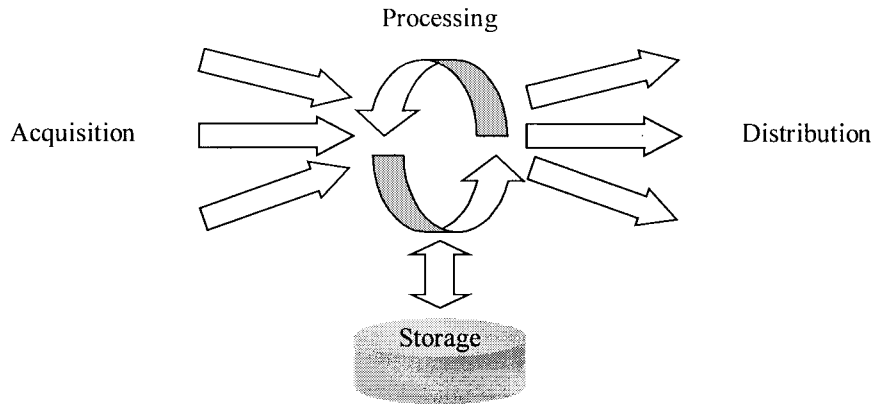


Fig. 1. Information flow model.

of analog content and are instead content based, what they represent, symbolize, and mean. By content based, we mean that such tools will enable ordinary users to act on the intellectual contents of images. Developing imaging tools further will be helpful to nearly everyone, but it will be essential to the future of technology in education. Students learn best through their own active efforts. Students need to control their visualization resources. The development of new media tools for education is becoming an urgent need as K–12 schools increase their dependence on digital multimedia information. Students conducting research on the Web and in other, proprietary digital archives need more effective means of retrieving relevant media objects; students working to make sense of information resources need better tools for annotating such objects; and students representing their ideas need more manageable tools for editing and manipulating media objects in the information production process. In a digital information environment, students of all ages can become more thoroughly engaged in the academic processes of information retrieval, analysis, and production, but they will need powerful, yet simple, information tools.

In this paper, we examine the state of the art, ongoing research, and the significant challenges that lie ahead in order to develop the next-generation techniques for content representation, creation, and searching. Our interest is not just on technology but rather on how audio-visual information can become part of our communications palette. We focus on education, as it is a particularly challenging domain for examining what it takes for new technical results to become an integral part of the users' expressive tools. The lessons learned will of course apply to any field of human endeavor.

II. State of the Art

To conceptualize multimedia systems, it is essential to have a model for describing the data flow in both traditional and new media systems. Education, and in fact any information-based application, shares the common work-flow tasks of acquiring, processing, storing, and distributing information, as shown in Fig. 1. Three aspects of this model distinguish traditional and new media: digitization, interactivity, and content awareness. The full digitization of information and networks allows for flexibility, easier integration, and immediate communication between the stages of the work-flow. Interactivity provides the user with the ability to affect the information flow immediately and enables processing of user input and feedback. Content awareness advances our traditional notion of audio-visual information to one consisting of objects rather than just image pixels or audio samples.

The vast majority of today's natural content is captured using traditional cameras and microphones. Even with current trends toward digitization and computer integration, this *acquisition* process fails to capture a significant part of the inherent structure of visual and aural information, a structure of tremendous value for applications. Various alternative visual sensing mechanisms have been proposed, including stereo capture and analysis, three-dimensional (3-D) scanners, and depth imaging, but require substantial hardware support and/or cannot operate in real time [58], [61], [77], [78]. Researchers at Columbia University, New York, recently introduced the OmniCamera, the first full-view video camera that allows users to select any viewpoint, without distortion and without any moving parts [78]. Such tools are not yet available, however, to regular users or content creators.

Acquisition is directly linked to digital *representation*. The emphasis in representation for the past several decades has been on compression: the description of a signal (audio, image, or video) with as few bits as possible [16], [31], [37], [59], [83]. When content becomes accessible via a computer, however, compression becomes just one of many desirable characteristics. Features such as object-based design, integration with synthetic (computer-generated) content, authentication, editing in the compressed domain, scalability and graceful degradation, flexibility in algorithm selection, and even downloadability of new tools are quickly becoming fundamental requirements for new media applications [50]–[52]. Several existing techniques [23], [101], [102] partially support some of these features, but the problem of an integrated approach is far from being solved. In addition, issues of compression efficiency for new forms of content (e.g., omnidirectional video) are open problems.

Storage facilities are required for large multimedia archives. However, it is still difficult to retrieve real-time information from distributed, heterogeneous sources. Research efforts have largely concentrated on the design of isolated sever components (e.g., file system or scheduling), without considering their interaction within the entire system. The emergence of multimedia on demand as a potential application for cable subscribers or Internet users has fueled research into ways to store and stream digital audio-visual information. There have been several efforts to build prototype systems (see [12], [13], [60], and references therein) and several trials for commercial services. Actual offerings, at least in the United States, have not been forthcoming, as the business issues do not yet seem to match the market's requirements. New issues in developing content-aware middleware and optimal resource allocation in networked distributed storage environments have also emerged as interesting topics.

Stored content needs to be easily accessible to users and applications. Current *retrieval* systems are unable to extract or *filter* information from multimedia sources on a conceptual level. The growing volume of audio-visual information makes it impossible to rely exclusively on manual text-based labeling techniques, especially if the type and organization of the information of interest is not known in advance. Work is being done today to automatically catalogue digital imagery in subject classes [18], [28], [29], [75], [93] or to recover high-level story structure from audio-visual streams [38], [71], [95], [106], but such efforts are ultimately limited by the rudimentary nature of the format of the source material and by the details of the representation techniques. Automated indexing based on simple low-level features such as color, texture, and shape has been to a large degree successful [2], [26], [81], [84]; we have developed several Web-based prototypes (e.g., VisualSEEk [92] and WebSEEk [93]) that demonstrate such capabilities integrated with spatial and text queries. Such low-level features, however, do not provide complete solutions for most users. Beyond simple queries for specific forms of information, users would like capabilities to extract information at higher levels of abstraction and track the evolution of concepts at higher semantic levels. Except for analysis of quantitative data archives (data mining), there has been far too little work on such high-level search concepts for multimedia information.

Processing of retrieved content by using computer-based *manipulation* tools is rapidly growing, even within the traditional film and television media. Lower cost editing and authoring suites (e.g., several commercial editing software on personal computers) bring these capabilities closer to regular users but with less flexibility and/or quality. Compressed-domain editing, which we have introduced in our WebClip prototype [71], [73], helps to reduce the quality degradation and increase flexibility and mobility. However, there is still a dichotomy between story-telling concepts (semantically meaningful entities, or objects) and the mechanisms used to manipulate these concepts on the chosen media (low-level pixels and audio samples).

The *creation* of high-quality structured multimedia content is an extremely laborious task and requires expensive specialized infrastructure and significant training. There is currently a large set of commercially available software packages addressing the creation of synthetic content (see [63] and [64] for a detailed description and comparison). These packages provide rather sophisticated capabilities for creating presentations, which often are capable of very sophisticated interaction between the user and content as well as between elements of the content itself. It is interesting to note that a notable few of these products follow an object-based design, treating the various—synthetic—content elements as individual objects.

Networking research for content *distribution* is currently fragmented, with researchers working on separate "native" asynchronous transfer mode (ATM), Internet, and mobile network models and prototypes. Development of ATM technology is driven by the ATM Forum, while the Internet Engineering Task Force (IETF) is responsible for Internet protocols. From modest experiments transmitting audio from IETF meetings in 1992, the Internet multicast backbone (MBone) and the associated set of Internet protocol (IP) audio/video multicast tools have seen widespread use. The recently launched Internet 2 project [42] addresses the foundation for a next-generation Internet infrastructure based on high-bandwidth connections, driven by the needs of future educational and research applications. Despite the different engineering approaches, the fundamental question yet to be successfully addressed is cost-effective quality of service provisioning. Even though we do not discuss networking issues in this paper, we should point out that they are of fundamental importance for successful development and deployment of new media applications.

In terms of educating engineering researchers and entrepreneurs that can successfully tackle these challenges, there are currently very few educational programs attempting to integrate media-related subjects in coherent cross-disciplinary curricula. At the same time, national policy in the United States is shifting the attention of K–12 educators away from teaching about computers toward teaching with them, integrating networked multimedia into each and every K–12 classroom. The U.S. Department of Education has issued a long-ranged plan, "Getting America's Students Ready for the 21st Century: Meeting the Technology Literacy Challenge." It calls on the country to meet four goals by the year 2000.

1) All teachers in the country should have the training and support they need to help students learn to use computers and the information superhighway.
2) All teachers and students should have access to modern multimedia computers in their classrooms.
3) Every classroom should be connected to the information superhighway.
4) Effective software and on-line learning resources should be an integral part of every school's curriculum.

The challenge to the engineering research community is to provide the know-how necessary to meet these goals on a global scale.

The focus in this paper is in the areas of representation, searching, creation/production, and editing. These capture the entire flow model except from storage and distribution. While the latter are of equal importance and form an integral part of a complete system, the key challenges for education and most other applications are in tasks where technology becomes the mediator between the user's and an application's perception of content.

III. New-Media Technology

Information technology has been a major economic force worldwide for a number of years, and there is every indication that it will continue to be so for several years. We are witnessing the transformation of our society from one focused on goods to one based on information. The extraordinary growth of the World Wide Web in just a few years demonstrates the need for, and benefit from, easy exchange of information on a global scale. Until now, audio and video in information technology had been treated as digitized versions of their analog equivalents. As a result, the use that they afforded to application developers and users was rather limited. In the following, we discuss current and emerging techniques to change the paradigm that drives the mechanisms with which users experience media, demonstrating the tremendous opportunities that lie ahead for research and development of novel applications. Our emphasis is on education applications, but similar (if not identical) arguments and technological solutions are applicable to any media-related endeavor.

A. Representation

In the beginning of this decade, there was a very important shift toward digitization of professional audio-visual content. Technological development allowed systems to be built that are capable of dealing with the very high bit rates and capacities required by real-time audio-visual information. Systems like the 4:2:2 D-1 digital videotape recorder are now commonplace in high-end digital studios, even though they have not yet supplanted their analog equivalents. These systems are very useful tools for professionals but do not directly affect end users, as their use is hidden deep within the professional studio.

1) Standards: A key development for the creation of services and applications that use digital delivery was the development of audio-visual compression standards. Following on the heels of the International Telecommunications Union—Telecommunications Sector (ITU-T) H.261 [56], [80] specification that addressed low-bit-rate video conferencing, the International Standards Organization (ISO) Motion Pictures Experts Group (MPEG) [37], [44], [45], [80] standards provided a solution that addressed the needs of the audio-visual content-creation industry [television (TV), film, etc.]. With these specifications, there was a solid common ground for the development of decoding hardware for end-user devices as well as encoding hardware for use by the content-development community. Interestingly, the increase in speed of general-purpose microprocessors within a period of just a few years now affords software decoders with real-time performance.

MPEG-1 [44] addresses compression for CD-ROM applications, with a combined bit rate of approximately 1.41 Mbps (single-speed CD-ROM). The target video signal resolution is one quarter that of regular TV (288×352 at 25 Hz or 240×352 at 30 Hz), with a coded rate of about 1.15 Mbps. The stereo audio signal has a frequency of 48 kHz, and using 16-bit samples is coded at 256 Kbps. In terms of perceived quality, with these rates, video is comparable to VHS tape, whereas audio achieves virtual transparency. MPEG-1 audio is used for digital audio broadcasting in Europe and Canada, while more than 2 million video CD players have been sold in 1996 in China alone. MPEG-1 has become one of the dominant formats on the Internet (coexisting with Apple's QuickTime and Microsoft's AVI), and virtually all graphics-card vendors today support MPEG-1 decoding either in software or in hardware. Also, Microsoft integrates a complete real-time software MPEG-1 decoder in its ActiveMovie software (a run-time version of which is included in all 32-bit Windows operating systems).

MPEG-2 [45] provided extensions in several important ways. It addressed compression of full-resolution TV and multichannel audio. It achieves studio quality at 6 Mbps and component quality at 9 Mbps; distribution quality with lower rates (e.g., 4 Mbps) is of course possible. It also addresses compression of high-definition (HD)TV and includes several optional scalability features. Since its introduction in 1994, MPEG-2 has allowed the creation of several digital content delivery services. In the United States, direct broadcast satellites have been deployed by various service providers (DirecTV, Primestar, Echostar, USB), offering more than 100 channels of MPEG-2 content directly to consumers' homes using very small (18-in) dishes. At the same time, the U.S. Federal Communications Commission has adopted a specification for HDTV terrestrial transmission building on the MPEG-2 specification (using Dolby AC-3 for audio), and the Digital Video Broadcasting Consortium is doing likewise in Europe. The recently introduced digital video disc or digital versatile disc will bring the convenience of audio CD-ROM's to video content.

These developments are very significant and represent important engineering milestones. They do not, however, fundamentally change the relationship between content producers and content consumers, where the consumer has a predominantly passive role. The same relationship is maintained, even if the delivery mechanisms and end-user devices are more sophisticated due to digitization. It is particularly interesting to examine the results of the use of computers within this digitized content environment.

2) Computers and Content Representation: The availability of low-cost encoding and decoding systems that resulted from the economies of scale (afforded by standardiza-

tion) allowed the creation of several low-cost tools that enhance regular computers with multimedia capabilities. For example, digital still and color cameras and interface boards can now directly capture images and video in Joint Photographic Experts Group (JPEG) and MPEG-1 formats (e.g., Hitachi's MP-EG1A), thus allowing one very easily to move raw compressed content into a computer. With the Internet providing a very low-cost distribution mechanism, consumer demand for such tools has been significant. At the same time, low-cost software tools became available to help in editing, such as Adobe Premiere. Still, we are not seeing any substantial increase in the use of audio-visual information despite the fact that users immediately embraced text and graphics on the Web, resulting in its astonishing growth. Users are being transformed from *information consumers* to *information producers*, but not of audio-visual content.

A basic problem is that raw content is seldom usable by itself. It requires painstaking manipulation so that the intended message is clearly conveyed. Content formats for distribution and editing are very different, however, especially for video. For example, MPEG-1 and MPEG-2 video cannot be easily edited due to the temporal dependency of the data. As a result, the processes of editing and acquisition for storage and distribution are not well integrated. Tools are being developed (described in Section III-C) to rectify these shortcomings. While these will go a long way in bringing audio-visual information closer to regular users, they still emulate analog processes of content creation. The reason is that the underlying representation of audio-visual information is directly bound to the analog acquisition process. Similar arguments hold for indexing and searching, where the problems are actually more pronounced due to the need to recover structure and semantics (see Section III-B).

3) The Need for a New Representation Framework: A fundamental limitation in our view of audio-visual information today is that it is extremely low level: composed of pixels or audio samples. This is the same as if we considered a text document as composed of the black-and-white dots that a printer produces. Clearly, our word-processing capabilities would not go very far if this were the level at which we had to operate each time we wanted to create a document. We have tools that completely abstract the details of printing and expose to us a world of semantically relevant entities: characters (of various fonts) that are combined to form words, sentences, and paragraphs. The tools themselves work behind the scenes to convert these characters into a form that can be printed or displayed on the screen. The user is free to focus on the actual content, ignoring the mechanics of its representation.

This is far from being the case for audio-visual information. Each time users want to create an audio-visual "document," they have to think and operate in terms of the constituent pixels or audio samples. Although a large number of tools are available to help in this process, there is a huge gap in the way we think about content and the way the tools are able to operate on it. There are two reasons for this shortcoming: 1) the use of audio-visual information in vertical applications and 2) preoccupation with bandwidth efficiency. Indeed, for the past several years, audio and video were only parts of complete systems, such as TV distribution or video conferencing. The behavior of the medium is not much different than regular analog TV, and the systems that host it are "closed." As a result, the only challenge facing engineering designers and researchers was to make the delivery of such content as cost effective as possible. Due to the cost of high-bandwidth connections, compression was the key design objective.

Compression, however, is only one aspect of representation. Our use of the term "representation" is indeed motivated by the fact that the way information is mapped into a series of bits can hold the key to several content attributes (coding, in this respect, is closer to representation than to compression). Requiring such mapping to be bit-efficient is just one of the possibilities; in the past, however, it had been considered as the only desirable one. There has been some slight change in perspective since the late 1980's, motivated by new types of communication channels. In particular, packet video (i.e., transport of compressed video over packet-based networks), wireless channels, etc., gave rise to issues of scalability and graceful degradation [80], [102]. It is interesting to note, however, that such features still address content-delivery issues and are only tangentially interesting for end users that want to do more than just see video being played back.

Our view, then, of media representation is much broader than compression. Several important media-engineering problems arise from the inadequacy of representation and could hence become obsolete by proper design of the way we put our visual and aural ideas into bits. By integrating the appropriate set of features, users as well as application programs would have the right "hooks" through which they can expose much richer sets of functionalities and ignore details that are of absolutely no interest to end users.

The key for braking this barrier lies in bridging the users' notion of semantically meaningful entities, with the elemental units dealt with in the representation framework. Currently, these units are samples or pixels, out of which pictures or picture sequences are built. From a user's point of view, though, what is important is the entities such sequences contain, what are their interrelationships, how they evolve over time, and how someone could interact with them. Following this reasoning, the notion of *objects* emerges quite naturally. These are audio-visual entities that have an independent nature in terms of the information they contain as well as the way they are represented. At the same time, they are something with which end users can relate, as they directly map story-telling concepts to groups of bits that can be manipulated independently.

There are several direct benefits of such an object-based approach. First, we allow the *structure of the content* to survive the processes of acquisition, editing, and distribution. This information is crucial in order to allow further editing or indexing and searching, since the difficult task of segmentation is completely eliminated. Today, this structure, which users painstakingly introduce during con-

tent creation, is completely eliminated by the distribution formats in popular use. Objects also allow the integration of natural and synthetic (computer-generated) content, each one represented in their native formats. In addition, they are natural units for user interaction. Last, but not least, compression of individual objects can be as efficient as one desires; in other words, compression efficiency does not have to be compromised because of the additional degree of flexibility.

4) The MPEG-4 Standard: We have been working within the ISO MPEG-4 standardization effort [2], [47], [50]–[52] in order to make such an object-based representation a universally available standard. MPEG-4 is the latest project of the MPEG group, being developed by more than 300 engineers from 20 countries around the world. It is currently in working draft status, and version 1.0 is scheduled to become an international standard in January 1999.

It will define tools with which to represent individual audio-visual objects, both natural and synthetic, as well as mechanisms for the description of their spatio-temporal location in the final scene to be presented to the user. The receiver then has the responsibility of composing the individual objects together for presentation. In the following, we briefly examine MPEG-4's features in more detail.

a) Visual object representation: MPEG-4 addresses the representation of natural visual objects in the range of 5 Kbps to 4 Mbps [54]. In addition to traditional "texture" coding, MPEG-4 specifies tools to perform shape coding. The combination of the two allows the description of arbitrary two-dimensional (2-D) visual objects in a scene. Both binary and "grayscale" alpha channel coding are currently considered. In addition, there are features for object scalability and error resilience. To a large extent, the algorithms used in MPEG-4 are quite similar to those employed in MPEG-2 and H.263. For still images, however, MPEG-4 is considering the use of zero-tree coding using wavelets [67], [88], since similar performance is achieved with other techniques but with the added benefit of scalability.

An important new direction in MPEG-4 is an effort to integrate natural and synthetic content, enabling synthetic–natural hybrid coding. In this respect, the visual component of the MPEG-4 specification addresses face-animation issues and has defined an elaborate set of face-animation parameters that can drive 3-D facial models. More traditional synthetic content such as text and graphics is, of course, included as well.

b) Audio object representation: Similarly, the audio component of the standard [53] addresses coding of single-channel audio at bit rates ranging from 2–64 Kbps and at higher bit rates for multichannel sources. The recently developed MPEG-2 advanced audio coding specification (a technique developed without the backward-compatibility requirement of MPEG-2 audio and hence achieving better performance) is included as well. Various forms of scalability are supported. In terms of synthetic content, basic musical-instrument digital interface and synthesized sound support is included, as well as speech synthesis from text and prosodic information.

c) Scene description: Scene description is defined in the systems part of the MPEG-4 specification [2], [55] and represents the most radical departure from previous MPEG efforts. It forms the glue with which individual objects are combined together to form a scene. The MPEG-4 scene description borrows several concepts from virtual reality modeling language (VRML) [1] (developed by the VRML Consortium but also an ISO draft international standard [46]). Scenes are described in a hierarchical fashion, forming a tree. Nodes within this tree either specify scene structure (e.g., spatial positioning, transparency, etc.) or denote media objects. Media-object nodes are associated with elementary streams using object descriptors, data structures carried separately from both the scene description and object data. This indirect association allows MPEG-4 content to be carried over a large variety of transport networks, including the Internet, ATM, or broadcast systems. For systems without proper multiplexing facilities, MPEG-4 defines its own multiplexing structure; its use, however, is optional.

Fig. 2 shows an overview of an MPEG-4 terminal. We use the term "terminal" in its most general sense, including both dedicated systems (e.g., set-top boxes) and programs running in a general-purpose computer. As is shown in the figure, the terminal receives individual objects as well as a description on how they should be combined together in space and time to form the final scene that will be presented to the user. It is up to the terminal to compose and render the objects for presentation. This essentially pushes the complicated task of composition from the production side all the way to the end-user side. This shift is critical for simplifying content creation, editing, and even indexing.

The types of operations allowed by scene-description nodes parallel the functionality of VRML nodes. In addition, interaction follows the same structure of event routing. The two approaches, however, are quite different in that MPEG-4 describes a highly dynamic scene that evolves over time based on external events (information being transmitted from the sender or obtained from a file), whereas VRML is addressing statically defined 3-D worlds that allow navigation. As a result, MPEG-4 scene descriptions can be updated dynamically, while the scene-description channel has its own clock reference and decoding time-stamps to ensure proper clock recovery and synchronization. In addition to this "parametric" scene description, an alternative "programmatic" methodology is also being considered. This is based on the use of the Java [33] language for controlling scene behavior. Programmatic control, however, does not extend to decoding or composition operations, thus avoiding performance limitations for such compute-intensive actions.

The World Wide Web Consortium has also initiated work in the specification of synchronized multimedia presentations in its Synchronized Multimedia (SYMM) working group [104]. This effort uses a textual format and does not address media representation, focusing only on the scene-

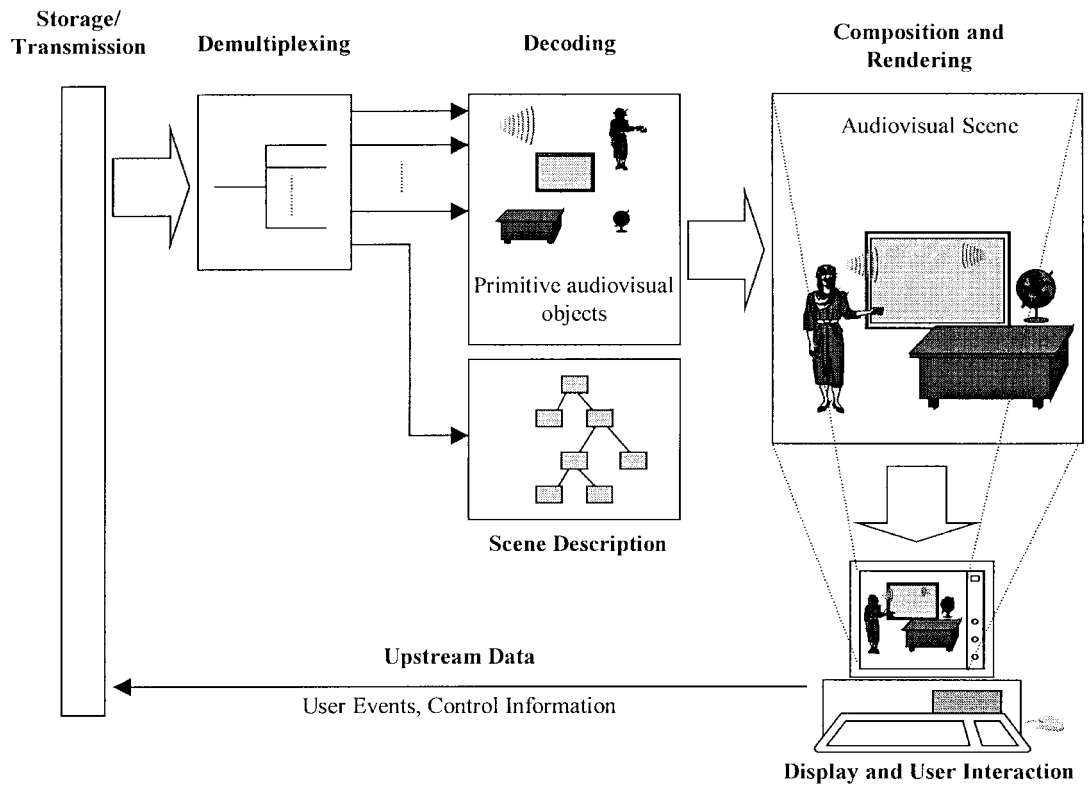


Fig. 2. Overview of an MPEG-4 terminal.

description aspects. As a result, it may not be able to provide the tight coupling between audio-visual objects desired in real-time audio-visual scene creation (for example, video would be treated as rectangular frames only).

Evidently, some overlap with other specifications is unavoidable considering the extensive scope that MPEG-4 has adopted. The challenge is to provide an integrated platform where both 2-D and 3-D, natural and synthetic, audio and visual objects can coexist and be used to create powerful and compelling content. For more information on MPEG-4, we refer the interested reader to several special issues of IEEE and EURASIP publications dedicated to the subject [40], [90], [91], as well as the official MPEG Web site [47].

5) Representation and Software Development: The power of objects can be fully utilized only with appropriate software-development tools. Indeed, in the 50-year history of media representation and compression, the lack of software tools is particularly striking. This has made the task of application developers much more difficult, as they have to become intricately familiar with the details of compression techniques.

The use of source coding, with its bit-oriented nature, directly conflicts with the byte-oriented structure of modern microprocessors and makes the task of handling coded audio-visual information more difficult. A simple example is fast decoding of variable length codes; every programmer that wishes to use information using entropy coding must hand-code the tables so that optimized execution can be achieved. General-purpose programming languages such as C++ and Java do not provide native facilities for coping with such data. Even though other facilities already exist for representing syntax (e.g., ASN.1-ISO International Standards 8824 and 8825), they cannot cope with the intricate complexities of source-coding operations (variable-length coding, etc.).

We are developing an object-oriented media-representation language intended for media-intensive applications called "Formal Language for Audio-Visual Object Representation" (Flavor) [19], [20], [23], [25]. It is designed as an extension of C++ and Java in which the type system is extended to incorporate bit-stream representation semantics (hence forming a syntactic description language). This allows the description, in a single place, of both the in-memory representation of data as well as their bit-stream-level (compressed) representation. Also, Flavor is a declarational language and does not include methods or functions. By building on languages widely used in multimedia application development, we can ensure seamless integration with an application's structure. Flavor is currently used in the MPEG-4 standardization activity to describe the bit-stream syntax. Fig. 3 shows a simple example of a Flavor representation. Note the presence of bit-stream representation information right after the type within the `class` declaration. The `map` declaration is the mechanism used in Flavor to introduce constant or variable-length code tables (1-to-n mappings); in this case, binary code words (denoted using the `0b` construct) are mapped to values of type `unsigned char`. Flavor also has a full complement of object-oriented features pertaining to bit-stream representation (e.g., "bitstream polymorphism") as well as flow-control instructions (`if`, `for`, `do-while`, etc.). The latter are placed within the declaration part of a

```
map SampleVLC(unsigned char) {
        0b0,  2,
        0b10, 5,
        0b11, 7
}
class HelloBits {
        int(8) size;
        int(size) value1;
        unsigned char(SampleVLC) value2;
}
```

Fig. 3. A simple example of flavor.

class, as they control the serialization of the class' variables into a bit stream.

We have developed a translator that automatically generates standard C++ and Java code from the Flavor source code [25], so that direct access to, and generation of, compressed information by application developers can be achieved with essentially zero programming. That way, a significant part of the work in developing a multimedia application (including encoders, decoders, content creation, and editing suites, indexing, and search engines) is eliminated. Object-based representations, coupled with powerful software development tools, is a critical component for unleashing the power of audio-visual information and making it available in a simple and intuitive form to regular users.

6) Algorithmic Content Representation: By extending the notion of object-based representation to include "programmatic" description of content, interesting new possibilities arise. By programmatic, we mean that content itself is described by a program instead of a series of bits that have a direct functional relationship to constituent pixels or audio samples. The proliferation of Java as a downloadable executable format has already demonstrated the power of downloadability. In the same way that useful application components can be downloaded when needed (and hence do not need to be provided in advance), a similar approach can be followed in content representation. For synthetic content, this can provide significantly advanced flexibility. As was mentioned in Section III-A4c, the approach is already considered for scene description within the MPEG-4 standardization activity.

This line of reasoning leads quite naturally to the consideration of a terminal as a Turing machine: the information transmitted to the receiver is not just data that will be converted to the original image or audio samples but also a program (possibly accompanied with data) that will be *executed* at the receiver to reproduce an approximation of the original content. Our traditional theoretical tools, based on information- and rate-distortion theories [7], [16], are not equipped to properly pose questions of efficiency in such a framework, as they completely ignore the internal structure of the receiver/decoder. Information theory asks the question: what is the smallest average number of bits needed to represent a given stochastic source. Rate-distortion theory addresses the same question but allows bounded distortion in the representation. Algorithmic description of information has long been addressed in traditional Kolmogorov

	Stochastic	Deterministic
Lossless	Entropy H(X)	Complexity K(x)
Lossy	Rate Distortion R(D)	Complexity Distortion C(D)

Fig. 4. Media-representation theories.

complexity theory [66], which addresses the question: what is the smallest length of a program, which, when ran in a Turing machine, will produce the desired object? This length is called the complexity of the particular object. It is interesting to note that this is not a stochastic measure but rather an inherent deterministic property of the object. It is a well-known result that for ergodic sources, complexity and entropy predict the same asymptotic bounds.

We are developing the foundations of a new theory for media representation called "complexity distortion theory." It combines the notions of objects and programmable decoders by merging traditional Kolmogorov complexity theory and rate-distortion theory by introducing distortion in complexity. We have already shown that the bounds predicted by the new theory for stochastic sources are identical to those provided by traditional rate-distortion theory [96], [97]. This completes the circle of deterministic and stochastic approaches for information representation by providing the means to analyze algorithmic representation where distortion is allowed. This circle is shown in Fig. 4. We are currently working toward practical applications of these results. Challenging questions of optimality or just efficiency in the presence of resource bounds (space or memory, and time) are of particular importance. In contrast to traditional theories, the use of such a framework allows us to pose such questions in a well-defined analytical framework, which may lead to promising new directions.

7) Key Research-and-Development Issues: To transcend our traditional pixel- or sample-based view of media, it is essential to incorporate in the digital representation of content as much of its original structure as possible. As the representation characteristics define, to a large extent, the possible operations that can be performed later on (indexing, searching, editing, etc.), the implications to the entire chain of media operations, from creation to distribution and playback, can be tremendous.

The following is a brief list of important technical barriers and research opportunities on issues that can greatly contribute toward this new viewpoint on representation:

- sensors that can capture 3-D information of content (e.g., depth—RGBD—cameras or omnidirectional cameras);
- real-time object segmentation tools for both visual and audio content;
- tools for encoding arbitrary objects, 2- or 3-D, both visual and aural;
- better understanding of the relationships between natural and synthetic content, seeking a common framework for the description of both;

- software tools for simplifying access to the internal characteristics of content by application developers;
- universally accepted standards for distributing object-based content;
- easy-to-use tools for enabling content creation by nonexpert users.

Parts of some of these issues are already being addressed. Even so, we expect that it will take several years before the fruits of this paradigm shift can be evidenced in the content-creation arsenal of regular users. Indeed, beyond making such technology available, its use requires thinking in modalities previously ignored. Although most people already have a quite rich subconscious visual and aural vocabulary due to film and television, its conscious use for personal communication is by no means a trivial change.

B. Searching

As various information sources prevail on-line, people have become more dependent on tools and systems for searching information. We search for content to explain ideas, illustrate concepts, and answer questions, all in the process of acquiring and creating knowledge. In the multimedia era, we tend to search for media-rich types of information, including text, graphics, images, videos, and audio. However, the utilities we use in content searching are still very primitive and far from satisfactory. The problem is particularly acute for visual content.

How does a student find an image or video clip from a large on-line encyclopedia that contains thousands of hours of historic video? How does a video journalist find a specific clip from a myriad of video tapes, ranging from historical to contemporary, from sports to humanities? Researchers in several disciplines such as image processing and computer vision, data base, and user interface are striving to provide solutions to finding visual content. In this section, we discuss various levels of content searching and different modalities of searching, present our experience in developing visual search engines, describe the general visual search system architecture, and discuss several important research issues in this area.

1) Different Search Levels:

a) Conceptual levels: People want to search for information based on concepts, independent of the media type of the content. A user may want to find images about "President Clinton discussing the budget deficit in a press conference" or images of "a suspension-style bridge similar to the Golden Gate Bridge." In the first example, we are concerned with the event, action, and place captured by the images, while in the second example, we are more interested in the concept conveyed by the images. The human vision system recognizes the image content at all levels, ranging from the high level of semantic meanings to the low level of visual objects and attributes contained in the images. But computers are still not able to achieve the same level of performance.

Image/video classification tries to fill the gap by linking the meanings of images to words. This requires a manual or, at best, semiautomatic process. Human operators need to decide what information to index, such as information in the categories of "who," "when," and "what." These data, called metadata, are extrinsic to the images and are used to describe the meanings of the images and videos. Selection and definition of metadata is not trivial. As discussed in [89], images have meanings at different levels, ranging from "preiconography" and "iconography" to "iconology." No manual assignment of image content descriptions will be complete. The choice of indexing information should depend on the intended use of the image collection. For example, medical-image domains and art/humanity domains clearly require different choices of indexing terms.

Several image archives, including Internet stock houses (e.g., Corbis, Yahoo) and archives at public institutes (e.g., The Library of Congress) are developing special taxonomies for cataloging visual content in their collections. But the lack of interoperable standards among customized cataloging systems will prevent users' seamless access to visual content from different sources. This problem calls for an important effort to standardize a core set of image subject classification schemes. Efforts such as the CNI/OCLC metadata core elements [103], the audio-visual program metadata work by EBU/SMPTE [18], and the MPEG-7 [48] international standardization effort have started to address issues along these lines.

b) Syntactic levels: Images and videos are composed of scenes and the spatio-temporal domain of visual objects, just like the real world captured by the images. Unlike the semantic meanings that require viewers' familiarity and knowledge of the subject, information in the syntactic level allows for image characterization by visual composition. At the syntactic level, we may want to find images that include the blue sky on top and an open green field of grass in the foreground, videos including a downhill skier with a zigzag motion trail, or a video clip containing a large, fast-moving object and a loud explosive sound track. Information at this level usually corresponds to low-level visual attributes of the objects in the images or videos. They are hard to index by using words due to the complexity and numerous aspects of the visual attributes. But automatic image/video analysis may provide promising solutions at this level. Searching for images by visual content provides a promising complementary direction with the text-based approach. The visual features of the images and video provide an objective description of their content, in contrast to the subjective nature of the human-assigned keywords. Furthermore, our experience indicates that integration of these two domains (textual and visual features) provides the most effective techniques for image searching.

In the area of content-based visual query, there has been substantial progress in developing powerful tools that allow users to specify image queries by giving examples, drawing sketches, selecting visual features (e.g., color, texture, and motion), and arranging the spatio-temporal structure of the features [4], [11], [26], [84], [92]. Usually, the greatest success of these approaches is achieved in specific domains, such as remote sensing and medical applications [65], [85]. This is partly due to the fact that in constrained domains,

it is easier to model the users' needs and to restrict the automated analysis of the images, such as to a finite set of objects. In unconstrained images, the set of known object classes is not available. Also, use of the image search systems varies greatly. Users may want to find the most similar images, find a general class of images of interest, quickly browse the image collection, and so on. We will compare different modalities of image searching in the following subsection.

2) Different Search Modalities: Images and videos contain a wealth of information and thus cannot be characterized easily with a simple indexing scheme. Many promising research systems have been developed by integrating multiple modalities of visual search [15], [34].

a) Text-based query: The use of comprehensive textual annotations provides one method for image and video search and retrieval. Today, text-based search techniques are the most direct and efficient methods for finding "unconstrained" images and video. Textual annotation is obtained by manual input, transcripts, captions, embedded text, or hyperlinked documents [9], [38], [76], [87], [99]. In these systems, keyword and full text searching may also be enhanced by natural language processing techniques to provide greater potential for categorizing and matching images. However, the approach using textual annotations is not sufficient for practical application. Manual annotations are often incomplete, biased by the users' knowledge, and may be inaccurate due to the ambiguity of the textual terms.

The integration of visual features and textual features provides promising avenues for cataloging visual information on-line, such as those used on the Internet. Our Web-based search engine, WebSEEk [93] explores this aspect and demonstrates significant performance improvement by using both the text key terms associated with the images and the visual features intrinsic to the images to index the vast amount of visual information on the Internet. We have found that the most effective method of searching for specific images of interest is to start with a keyword search or subject browsing and then follow up with a search based on visual features, such as color.

b) Subject navigation: Images and videos in a large archive are usually categorized into distinctive subject areas, such as sports, transportation, lifestyle, etc. An effective method in managing a large collection is to allow for flexible navigation in the subject hierarchy. Subject browsing is usually the most popular operation among leisure users. It is often followed by more detailed queries once the users find a specific subject of interest.

A balance between the depth and width of the subject hierarchy should be maintained. A deep division of subjects may make it difficult for users efficiently to select the initial browsing path. On the other hand, broad definitions of subject areas may undermine the discrimination power of subject division. In addition, the order of subject levels in the subject hierarchy will also affect the users' ability to find the right target subject.

Usually, the subject hierarchy is developed in a way similar to that of top-down tree growing. But each image or video in the data base may be linked to multiple subjects in different levels. Fig. 5 shows the first level of subject hierarchy (i.e., taxonomy) in WebSEEk. The WebSEEk taxonomy contains more than 2000 classes and uses a multilevel hierarchy. It is constructed semiautomatically in that, initially, human assistance is required in the design of the basic classes and their hierarchy. Then, periodically, additional candidate classes are suggested by the computer and are verified with human assistance.

Classification of new images into the taxonomy is done automatically by comparing the associated key terms of images to the words describing each subject node. The performance in classifying visual information from the Web is quite good. We have found that WebSEEk's classification system provides over 90% accuracy in assigning images and videos to semantic classes. As mentioned earlier, however, each image may have semantic meanings in different aspects and at different levels. Using the associated terms from the associated hypertext mark-up language (HTML) documents and the file names will clearly not be sufficient to capture all of the various meanings of an image.

c) Interactive browsing: Leisure users may not have specific ideas about images or videos that they want to find. In this case, an efficient interactive browsing interface is very important. Image icons, video moving icons and key frames, and multiresolution representation of images are useful in providing a quick mechanism for users to visualize the vast amount of images or videos in the archive. A sequential, exhaustive browsing of each image in the archive is impractical. One approach is to use clustering techniques or connected graphs [108]. The former organize visually similar images in the same cluster (e.g., high-motion scenes, panning scenes). The latter link image nodes in the high-dimensional feature space according to their feature similarity. Users may navigate through the entire feature space by following the links from a node to any other neighboring nodes. The objective is for users to effectively visit any node in the entire image space by simple iterative browsing.

d) Visual navigation and summarization: Document summarization is a popular technique used in today's document search engines. It provides briefing about the content in one single document or multiple documents. The same concept can be applied to the visual domain. In the simplest form, a size-reduced image representation (e.g., an icon) can be considered as a visual summarization of the image. For video, the task is more challenging. Most systems segment the video into separate shots and then extract the key frames from each shot. A hierarchical key-frame interface or a scene-based transition graph is used as an efficient tool for users quickly to view the visual content from a long video sequence [71], [95], [106].

Another approach uses motion stabilization techniques to construct the background image from a video sequence and simultaneously track moving objects in the foreground [43]. The objects in the foreground and their motion trails can be overlaid on top of the mosaic image in the background to summarize the visual content in a video sequence. By

Fig. 5. Subject browsing interface for an Internet image search engine (WebSEEk).

looking at the mosaic summarization, users can quickly apprehend the visual composition in the spatio-temporal dimension. This technique is particularly useful for surveillance video, in which abrupt motions may indicate important events.

e) Search by example: Searching for images by examples or templates is probably the most classical method of image search, especially in the domains of remote sensing and manufacturing. Users use an interactive graphic interface to select an image of interest, highlight image regions, and specify the criteria needed to match the selected image template. The matching criteria may be based on intensity correlation or modified forms of correlation between the template image and the target images. Although correlation is a very direct measurement of the similarity between the template and the target images, this technique suffers from sensitivity to noises, sensitivity to imaging conditions, and the restrictive need of an image template.

f) Search by features and sketches: Feature-based visual query provides a complementary direction to the above search method using templates. Users may select an image template and ask the computer to find similar images according to specified features such as color, texture, shape, motion, and spatio-temporal structures of image regions [4], [26], [84], [92]. Some systems also provide advanced graphic tools for users to directly draw visual sketches to describe the images or videos they envision [11], [57], [92]. Users are also allowed to specify different weightings for different features. Fig. 6 shows an example of using a visual sketch describing the object color and the motion trail to find a video clip of a downhill skier.

The success of feature-based visual query relies on the fast response of visual queries and informative query results to let users know how the query results are formed and, perhaps, which feature is more important in determining the final query results. Ease of use is also a critical issue in designing such a query user interface. Our experience indicates that users are usually much less enthusiastic about this query method than others previously mentioned when the query interface is complex.

g) Search with agent software: Last, a high-level search gateway (or so-called metasearch engines) can be used to hide from users the complex details of the increasing number of search tools and information sources.

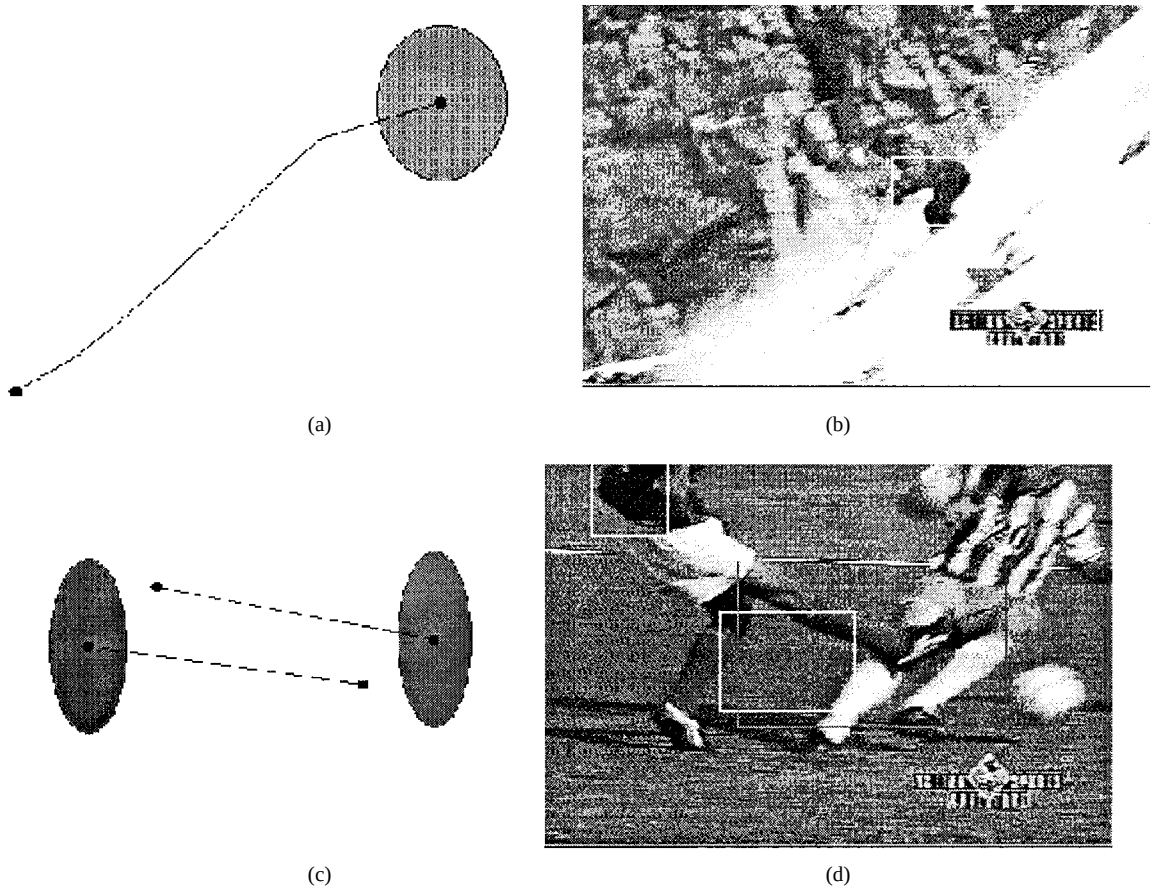


Fig. 6. Sketch-based visual queries and the returned videos. (Videos courtesy of Action Sports Adventure, Inc., and Hot Shots/Cool Cuts, Inc.)

The metasearch engine translates the user-specified query to forms compatible with individual target search engines, collects and merges the query results from various sources, and monitors the performance of each query and recommends the best target search engines for subsequent queries [17]. In the visual domain, the metasearch engines are in an early stage of development and will require substantial efforts in solving critical technical issues, such as performance evaluation and interoperable visual features [6].

3) System Architecture of Multimedia Search System: The general system architecture for a content-based visual search system is depicted in Fig. 7. We discuss the major components in the following sections.

a) Image analysis and feature extraction: Analysis of images and feature extraction plays an important role in both off-line and on-line processes. Although today's computer vision systems cannot recognize high-level objects in unconstrained images, low-level visual features can be used to partially characterize image content. These features also provide a potential basis for abstracting the semantic content of the image. The extraction of local region features (such as color, texture, face, contour, motion) and their spatial/temporal relationships is being achieved with success. We argue that the automated segmentation of images/video objects does not need to accurately identify real-world objects contained in the images. Our goal is to extract the "salient" visual features and index them with efficient data structures for fast and powerful querying. Semiautomated region-extraction processes and use of domain knowledge may further improve the extraction process.

b) Interaction loop including users: One unique aspect of image search systems is the active role played by users. By modeling the users and learning from them in the search process, image search systems can better adapt to the users' subjectivity. In this way, we can adjust the search system to the fact that the perception of the image content varies between individuals or over time. User interaction with the system includes on-line query, image annotation, and feedback to individual queries, as well as overall system performance. Image query is a multiiteration, interactive process, not a single-step task. Iterated navigation and query refinement is an essential key in finding images. Relevance feedback has been successfully used to adapt the weightings of different visual features and distance functions in matching images [39], [86].

User interaction is also useful in breaking the barrier of decoding semantic content in images. Learning through user interaction has been used in video browsing systems to dynamically select the optimal groupings of features for representing various semantic classes for different users at different times [74]. Some systems learn from the users' input as to how the low-level visual features are to be used in the matching of images at the semantic level. Un-

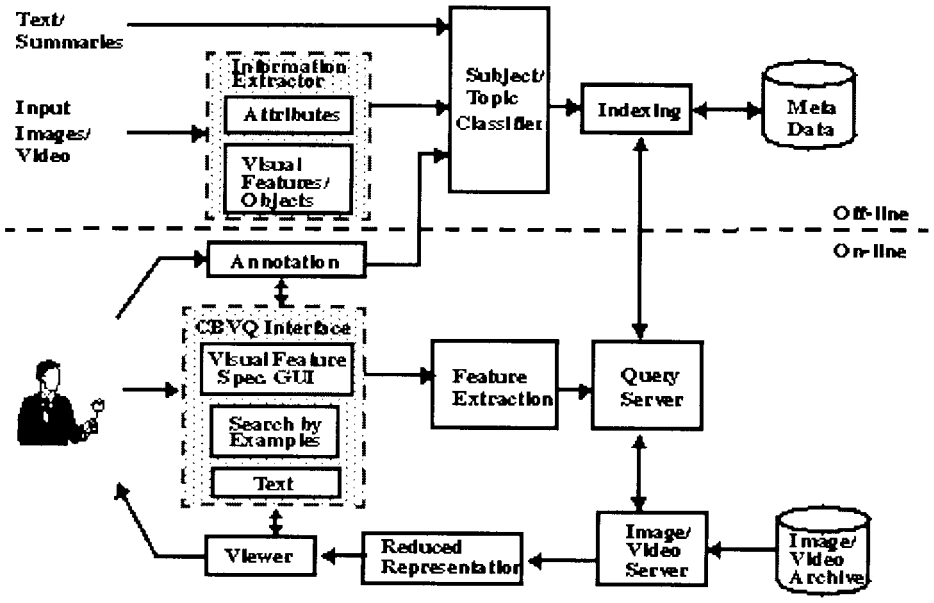


Fig. 7. A general architecture for content-based visual search systems.

known incoming images are classified into specific semantic classes (e.g., people and animals) by detecting predefined image regions and verifying spatial constraints [28].

c) Integration of multimedia features: Exploring the association of visual features with other multimedia features, such as text, speech, and audio, provides another potentially fruitful direction. Our experience indicates that it is more difficult to characterize the visual content of still images compared to video. Video often has text transcripts and audio that may also be analyzed, indexed, and searched. Also, images on the World Wide Web typically have text associated with them. In this domain, the use of all potential multimedia features enhances image-retrieval performance.

d) Efficient data base indexing: Visual features are extracted off-line and stored as metadata in the data base. Content-based visual query poses a challenging issue in that the variety and dimension of visual features are both very high. Traditional data base indexing schemes such as Kd trees and R trees [5], [30], [35], [98] cannot be directly applied to cases with such high dimensions. Most systems use techniques related to prefiltering to eliminate unlikely candidates in the initial stage and to compute the distance of sophisticated features on a reduced set of images [22], [36]. However, generalization of these techniques needs to be further studied in order to handle different types of distance metrics.

4) Key Research-and-Development Issues: Image/video searching requires multidisciplinary research and validation in real applications. Different research communities may focus on separate subareas, but an essential step in achieving a functional practical system is the participation of user groups in system development and evaluation. A real application like the high school multimedia curriculum (e.g., the Eiffel project, described later) can be used to establish an ideal testbed for evaluating the various research components discussed above.

In addition to a real application testbed, the following includes a partial list of critical research issues in this area (see [15] for more discussion):

- multimedia content analysis and feature extraction;
- efficient indexing techniques and query optimization;
- integration of multimedia;
- automatic recognition of semantic content;
- visual data summarization and mining;
- interoperable metadata standard;
- evaluation and benchmarking procedure;
- on-line information filtering;
- effective feature extraction in the compressed domain.

Some of these issues may have been active research subjects in other existing fields. But content-based visual search poses many new challenges and requires cross-disciplinary collaborative efforts.

C. Creation/Production

Audio-visual content creation today is a difficult, time-consuming task and requires significant expertise when high quality is desired. This is acceptable when producing a television program or a movie, where the value and return on investment are well defined, but not for regular computer users that want to venture into the realm of audio-visual content creation and communication. Educational environments in this sense are even more demanding, as it is important to make technology virtually transparent to potentially very young users while still keeping the cost at very low levels.

There has been extensive work over a number of years on the creation of synthetic content. Indeed, the entire field of computer graphics is essentially addressing synthetic content creation. This includes both 2-D and 3-D modeling, rendering, and animation, as well as graphical user inter-

faces, etc. (see [27] and [32] and references therein). The area is extremely mature and in recent years has been an indispensable component of professional content developers, especially in the movie industry (special effects, etc.).

We can in general identify three major categories of content-creation tools.

1) Authoring tools for synthetic content that come with proprietary players. This includes many commercial software packages available on PC and workstations today (see also Section II).
2) Authoring tools using *de jure* or *de facto* distribution standards for which several players are available. Here, the emphasis is on the distribution format. Key examples are VRML and HTML. The latter in particular can be considered as the text-based glue that provides a mechanism for combining components together.
3) Content-creation tools that are intended for image or video sequence synthesis. These are not concerned with playback capabilities and instead rely on external mechanisms for integrating content in traditional delivery mechanisms (MPEG, analog tapes, etc.). Some systems, however, are built for specific representation standards (e.g., motion JPEG, MPEG).

The first category is the one closest to the level of integration required by audio-visual content, but it addresses synthetic content and relies on proprietary formats for distribution and playback. The use of synthetic content relaxes some of the engineering design requirements, and in particular those of synchronization. In addition, these formats are not intended for so-called "streaming" or continuous delivery. In some cases, additional tools are provided for conversion to a format amenable to streaming (such as Enliven by Narrative Communications, which converts Macromedia Director files). This category is most popular in educational applications but is also dominant in corporate training and general CD-ROM title development.

The second category does not satisfy the requirements of audio-visual content creation. VRML, as we discuss in Section III-A, does not meet the dynamics of audio-visual content (e.g., handling a "feed" from a national broadcaster). HTML, even though it has been instrumental as a common denominator for exchanging documents that include text and graphics and has been the propelling engine of the Web, is a primarily textual facility. GIF animation certainly adds a dynamic flavor to the content, but still the primary message-bearing component is the text.

Last, the third category includes extremely powerful systems, but typically at significant cost and with the need for additional tools for preparing a finished product. Systems without special equipment usually compromise the performance significantly. This category is the one predominantly exposing a visual domain for content creation. An important issue for such tools is the requirements imposed on users in terms of additional equipment, software, and storage capacity. For example, generation of uncompressed frames results in the need for about 20 MB per second of content. Two minutes of such content can easily fill the entire disk of an average personal computer. Note also that additional disk space is needed for intermediate results or alternate versions; hence, increases in storage capacities will not necessarily solve this problem, even if they render it less acute. In addition, without special hardware, the speed performance is usually quite slow.

1) A New Object-Oriented Platform for Content Creation: In all of the above three categories of content creation, natural content is absent; at best, it is present as simple rectangular video windows. It is interesting to note that, to our knowledge, there have not been any research or development efforts addressing the needs of regular users for both synthetic and natural content-creation tools. We believe that this is exactly because of the limitations of today's frame-oriented, pixel-based representation, which leaves no other alternatives to application developers. As a result, the expressive power of imagery is not fully tapped. The MPEG-4 standard (see Section III-A4) provides a new object-oriented content-based framework and can be instrumental in this case in terms of providing a rich representation framework on which content-creation tools can be built.

Although segmentation of video objects from natural videos is still an open research issue, the authoring tools should take advantage of this synergistic framework and provide flexible manipulation at the object level. Video objects can be linked to the semantic concepts more directly than the restricted structures using frames or shots. For example, students may want to cut out one foreground object from one video sequence and experiment with combinations of different backgrounds in learning the aesthetic aspects of video shooting or film making. They may also want to create hyperlinks for video objects to link them to associated documents.

In collaboration with the Institute for Learning Technologies, and via the Eiffel project (see Section IV), we are examining the requirements for such content-creation tools for K–12 educators and students. We are developing a content-creation software suite called Zest, with which we will explore how the new object-based paradigm can unleash the power of audio-visual information for regular users. Using preexisting audio-visual objects or building ones from scratch, users have the flexibility to define the objects' spatial and temporal positioning as well as behavior. Creating appealing and rich content becomes a point-and-click operation on a spatial and temporal canvas. The created content is stored in the MPEG-4 format, and thus playback capability on various platforms will soon be available. We place special emphasis on simplicity and effectiveness rather than supporting a huge array of features (most of which typical users tend to underutilize). By testing our work in as demanding environments as K–12 schools, we believe that significant insight can be obtained so that the end result satisfies not only the needs of a technology-enabled curriculum but the broadest spectrum of end users as well.

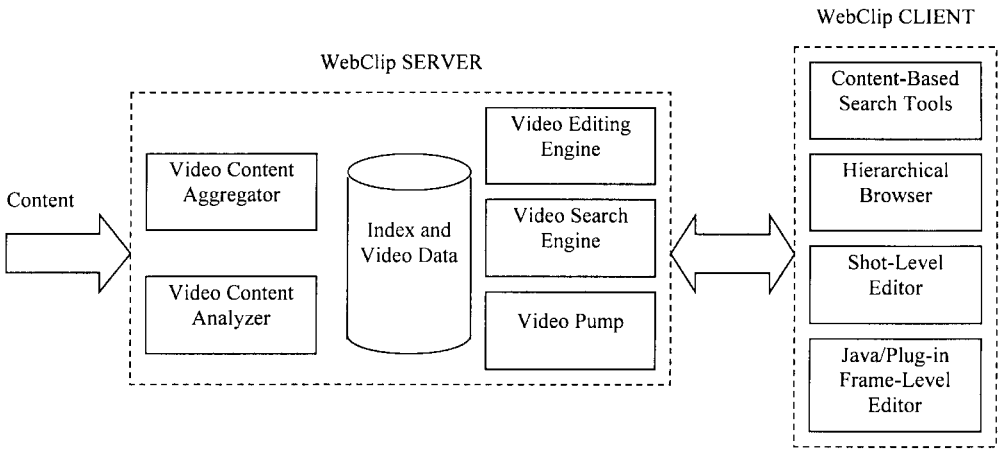


Fig. 8. Major components of a networked video editor, WebClip.

2) Content Creation in Distributed Networked Environments: Another dimension for enhancing multimedia content creation/production is to extend the authoring platform from stand-alone stations to distributed environments and from single-author systems to collaborative systems. In addition, ideal content-creation tools should allow users to manipulate content with the maximum flexibility in any medium they prefer (e.g., edit by video, edit by text, or edit by audio), on any level (including semantic) without distraction by the technical details, and at any location without significant difference in performance.

The above requirements have a profound technical impact on the development of advanced content-creation systems (particularly for video). First, they need to be responsive. User interfaces should have great interactivity and near real-time response. This is particularly important when dealing with young students in order to keep up with their attention span. Second, due to the massive size of multimedia, different levels of resolutions (in space, time, and content) should be provided. Multiresolution stages can be used to trade off content quality with requirements of computing/communication resources in real-time applications. Last, synchronization and binding among multiple media should also be emphasized so that editing can be easily done in any media channel.

a) A Web-based networked video editor: We present a networked video editing prototype, WebClip, to illustrate the above requirements and design principles. WebClip is a complete working prototype for editing/browsing MPEG-1 and MPEG-2 compressed video over the World Wide Web [71]–[73]. It uses a general system architecture to store, retrieve, and edit MPEG-1 or MPEG-2 compressed video over the network. It emphasizes a distributed network support architecture. It also uses unique compressed video editing, parsing, and search technologies described in [71].

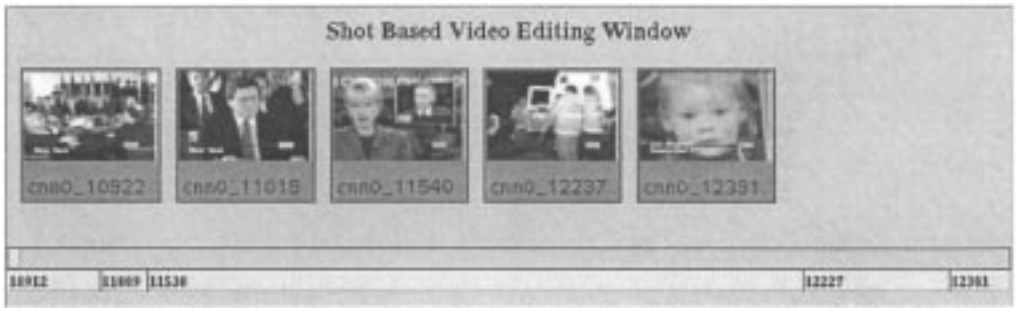
Other unique features of WebClip include compressed-domain video editing, content-based video retrieval, and multiresolution access. The compressed-domain approach [10], [14], [100] has great synergy with the network editing environment, in which compressed video sources are retrieved and edited to produce new video content, which is also represented in compressed form.

Major components of WebClip are depicted in Fig. 8. The video content aggregator collects video sources online from distributed sites. Both automatic and manual mechanisms can be used for collecting video content. The automatic methods use software agents that travel over the Web, detect/identify video sources, and download video content for further processing. The video-content analyzer includes components for automatic extraction of visual features from compressed MPEG videos. Video features and stream data are stored in the server data base with efficient indexing structures. The editing engine and the search engine include programs for rendering special effects and processing queries requested by users.

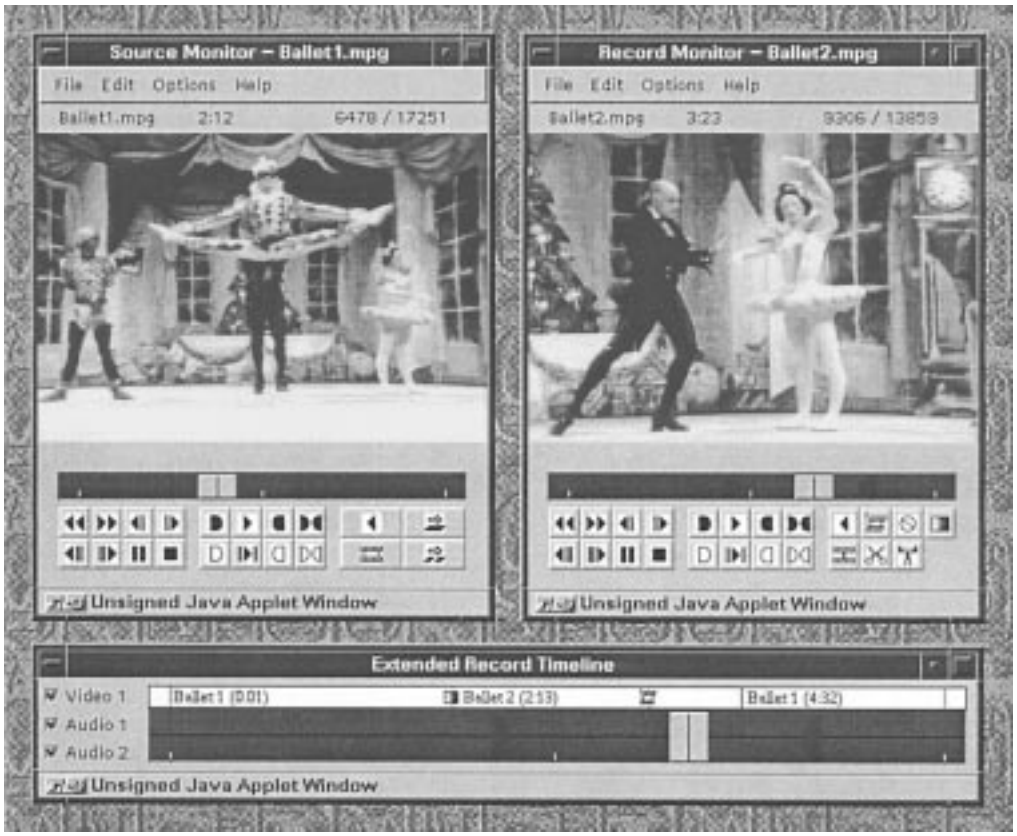
On the client side, the content-based video search tools allow for formulation of a video query directly using video features and objects. The hierarchical browser allows for rapid visualization of important video content in video sequences. The shot-level editor includes tools and interfaces for performing fast, initial video editing, while the frame-level editor provides efficient Java-based tools for inserting basic editing functions and special effects at arbitrary frame locations. To achieve portability, current implementations also include client interfaces written in Java, C, and a Netscape plug-in.

The frame- and shot-level editors are shown in Fig. 9. This multilevel editing design applies the multiresolution strategy mentioned above. The idea is to preserve the highest level of interactivity and responsiveness in any arbitrary editing platform. The shot-level editor is intended for platforms with low bandwidth and computing power, such as light-weight computers or notebooks with Internet access. The frame-level editor includes sophisticated special effects, such as dissolve, motion effects, and cropping. It is intended for high-end workstations with high communication bandwidth and computation power.

Before the editing process is started, users usually need to browse through or search for videos of interest. Various search methods discussed in Section III-B can be used for this purpose. In addition, WebClip provides a hierarchical video browser allowing for efficient content preview. A top-down hierarchical clustering process is used to group related video segments into the clusters according to their



(a)



(b)

Fig. 9. Multiresolution editing stages. (a) The shot-level editing interface. (b) The frame-level editing interface.

visual similarity, semantic relations, or temporal orders. For example, in the news domain, icons of key frames of video shots belonging to the same story can be clustered together. Then, users may quickly view the clusters at different levels of the hierarchical tree. Upper level nodes in the tree represent a news story or group of stories, while the terminal nodes of the tree correspond to individual video shots.

The networked editing environment, which takes compressed video input and produces compressed video output, also makes the compressed-domain approach very desirable. The editing engine of WebClip uses compressed-domain algorithms to create video cuts and special effects, such as dissolve, motion, and masking. Compressed-domain algorithms do not require full decoding of the compressed video input and thus provide great potential in achieving significant performance speedup [10]. However, existing video-compression standards, such as H.263 and MPEG, use restricted syntax (such as the block structure and interframe dependence) and may require substantial overheads for some sophisticated video-editing functions such as image warping.

3) Key Research-and-Development Issues: We envision a next-generation content-creation paradigm in which video content consists of either natural or synthetic objects from different locations, either live or stored. For example, video objects of a video program may not be stored in the same storage system. This type of distributed content is not unusual in on-line hypertext. Considering today's video-capturing methods, it may still be early to anticipate extensive use of this type of distributed video content. However, it may become popular in the future, as more video content will be created by using video editing tools

like WebClip and Zest and by reusing existing video from distributed sources. In such a distributed object-based video paradigm, video editors will need to handle new challenges related to synchronization, particularly for on-line real-time editing systems. Earlier work [68] on spatio-temporal composition of multimedia streams has addressed the issues with a higher granularity (e.g., video clip, audio sequence, text, and images) rather than at the arbitrarily shaped video object level. New research on object-level editing with support of real-time interactivity will be required.

IV. New-Media Applications in Education

Printed media have dominated education, making it bookish. This dominance arose not from some perverse error. It came about largely because experience captured in writing and reproduced through printing became effectively searchable, accessible to diverse persons at many locations over extended times through random access. Historically, this privileged written resources, making the experience they recorded transmutable far more easily into knowledge transmittable through formal education from one generation to another. The search modalities for new media collections described above begin to endow visual and auditory resources with the same sort of on-demand retrievability long enjoyed by printed resources. We plan to introduce these search modalities into classrooms and to help teachers and students apply them in the course of their work. These search tools can show up in a variety of educational applications. In the same way that educators have developed numerous strategies to teach writing, so will they develop ways to use these modalities to teach seeing.

The Columbia New Media Technology Center and the Institute for Learning Technologies at Teachers College have teamed together to work over the long term to pioneer these educational innovations and to engineer the digital media systems that can make them feasible. A five-year, multimillion-dollar U.S. Department of Education Challenge Grant for Technology in Education provides the core resources for the educational work, which will link 70–100 New York City public schools (more than 30 000 students) in a high-speed testbed for new curriculum designs [41]. Researchers developing the projects described above will work with students and teachers in participating schools. We are working to design classroom applications that take advantage of the content-based developments in representation, searching, and editing. These functions, crucial to advancing the state of the art technologically, also pertain directly to achieving major advances in the quality of education.

A. Representation

Educators seeking to integrate the use of information and communications technologies into mainstream classroom experience must ensure that those tools do not become the object of students' inquiries, but rather that they serve as a transparent means through which students study and learn *through* the technology. With traditional multimedia systems, the technology tends to get in the way of good learning. One of two things tends to happen. If the system is configured as a tool that students should use in an open-ended way of expressing their ideas and understanding, the technology often displaces the object of study, forcing the student to attend to it in order to do anything with the system. This results in the complaint that too much educational software is difficult to use. If, instead, the system is configured to convey information and to exercise students in recall and manipulation, it degrades the quality of interaction into structured multiple choices. This results in the objection that many programs accentuate a drill-and-practice mentality that bores and alienates students. In the one case, the act of representing a concept is too difficult; in the other, the complex act of representation becomes simplified into one of mere identification.

Consider, in contrast, the learning situation that becomes feasible with the object-based content representation tools described in Section III-A. It will become possible to develop a variety of learning resources in which students receive a set of primitive audio-visual objects and scene-description tools, which they then can use to construct representations of difficult concepts. With a relatively simple set of graphic primitives and scene descriptors, students could construct representations of cell mitosis or changing balance-of-power relations in nineteenth-century European history. The learning will focus only incidentally on the technology and substantively on the conceptual question at hand. The quality of interaction will, however, be rich and intense, for the students will need to create, not merely identify, the conceptual representation. Working in the context of our testbed of schools, we are developing teams of teachers, curriculum specialists, and engineers to identify important concepts that students can master by creating relevant representations using content-based imaging tools.

B. Searching

Whereas representation exercises are likely to become a technique in helping students master existing components of the curriculum, searching with content-based imaging tools will itself become an important element of the overall curriculum. The stock of human knowledge is rapidly going on-line. Developing skill at finding information and intellectual resources has been an important school curriculum goal for students going on to higher education, and a secondary objective in general education. In an information society in which the cultural assets of the civilization increasingly become available on-line to any person from any place at any time, the ability to select and retrieve those resources most pertinent to one's purposes becomes an increasingly important educational objective for all. Furthermore, as the stock of knowledge available on demand becomes increasingly a multimedia stock, content-based image search and retrieval grows in importance. Consequently we plan to concentrate considerable effort on developing the educational uses of such search tools.

Looking at content-based search tools as educators, we anticipate two major lines of development. One aims at developing the capacity of students to think visually and to

devise effective search heuristics with these resources. The other seeks to deploy the tools against important image resources available on the Web to improve educational experiences in subject-matter areas such as science and history. In both cases, content-based search tools discussed in Section III-B will provide tools that will enable us to work on both lines of development.

We anticipate that content-based search tools will allow educators to address the heuristics of visual thinking across a wide range of developmental stages. For instance, one could pose an interesting challenge to younger children, developing their capacity to think about the identifying characteristics of different animals by asking them to do a search that would retrieve pictures of giraffes—in side view and in frontal view. At a much further stage of educational development, one might challenge science students to develop a visual search of moving images that would return clips illustrating gravitational acceleration on falling bodies. Given a powerful set of content-based search tools, the range of queries that might be asked of our stock of images is limitless, and an important educational goal will be to develop the acuity with which students can form and pose such queries.

Building up students' capacities to pose effective queries with content-based retrieval tools will in turn make those tools a powerful source of substantive learning in a variety of fields. A picture is worth a thousand words, the saying goes. Yet education and the production of knowledge has remained largely verbal, not visual, because our storage and retrieval systems have been so exclusively verbal. With content-based search and retrieval tools, educators working at all levels face an interesting opportunity, finding ways to make the stock of images work as primary communicators of human thought and understanding. Through our school-based testbed, we are initiating sustained development efforts to develop these applications of content-based tools to the acquisition of knowledge across all levels of education. An initial fruit of such efforts is the "*Where Are We*?" program, which develops children's ability to use maps effectively [62].

C. Creation/Production

Interactive multimedia are often proclaimed to be a powerful force for educational improvement. In thinking about the educational uses of multimedia, we often pay too little attention to the question of who will create and manage its production. Elaborate productions designed far from the working classroom have the ironic effect of putting both teacher and student in a predominantly passive, responsive role. Interactive multimedia is much more significant when teachers and students have control over its production and can use it as a tool of communication, expressing their ideas and understanding of a subject. For this to happen, production tools need to be simple, powerful, and accessible.

As a result of the World Wide Web, a great deal of content in diverse media is becoming available to teachers and students. Traditionally, educators have seemed to face a difficult dilemma. On the one hand, in order to make education intellectually rigorous and demanding, they must impose a standardized regimen on students that alienates many. On the other, to engage each student in learning that he personally relates to, they must use projects that often become superficial and dubious in intellectual value. If students can build projects from the wealth of materials available on the Web, having control over their construction on the one hand but having to engage with the full scope of intellectual resources pertinent to those projects on the other, then possibilities for a pedagogy that attains exemplary intellectual breadth and rigor, while proving deeply engaging to the student, may be feasible [8], [69]. WebClip and Zest, discussed in Section IV-C, should prove to be very useful enabling software in implementing such a pedagogy. The Institute for Learning Technologies has extensive experience through the Dalton Technology Project in developing educational prototypes in which students create multimedia essays with multimedia resources over a local-area network [70]. Using new content-creation tools in testbed schools, we will reengineer such prototypes for use over the World Wide Web in a much wider educational setting.

In sum, engineers and educators share an essential design problem. The systems characteristics to be developed in creating content-based new media tools are precisely the functional characteristics that will make these tools educationally significant. Content-based new media are tools that will facilitate the production and dissemination of knowledge. And insofar as new media become tools for the production and dissemination of knowledge, they become powerful agents altering what is feasible throughout education. We expect technology advances will steadily empower a series of educational innovations, and efforts to implement those innovations will enable us to ready the technology for broad popular use.

V. Concluding Remarks

Next-generation new-media applications will start enabling people to use audio and visual resources in flexible, reflective ways. The long-term cultural implications of these developments are likely to be very significant. To move vigorously toward their realization we need to overcome key technical barriers, among them:

- the inability of existing sensors to capture the full view, complete structure, and precise identity of objects;
- the inability to directly extract information about content using existing techniques for multimedia representation and retrieval;
- the difficulty of providing easy-to-use techniques for analyzing, presenting, and interacting with massive amounts of information;
- lack of integration of existing networking models (IP, ATM, wireless) where none alone is capable of fulfilling all new-media application requirements, including ease of service creation, resource allocation, quality of service, and mobility.

In addition, we need to bring next-generation new-media applications into everyday use in a wide range of situa-

tions, preeminently in education. To make that happen, we will need to accomplish four things consistently, with all students, under all conditions:

- pose powerful generative questions in cooperative settings;
- end limitations on the intellectual resources available to students in their classrooms and in their homes;
- enable teachers and students to communicate beyond the classroom, as they want, around the world;
- provide advanced tools of analysis, synthesis, and simulation.

Effective application of next-generation content representation, creation, and searching, as discussed in this paper, will be an essential part in overcoming these technical barriers and making fundamental educational reform feasible under conditions of everyday practice.

ACKNOWLEDGMENT

The authors wish to thank the many graduate students who have contributed a great deal to the multimedia searching and editing work described in this paper over a period of many years, including J. R. Smith, J. H. Meng, W. Chen, H. Sundaram, D. Zhong, A. Benitez, and M. Beigi.

REFERENCES

[1] A. L. Ames, D. R. Nadeau, and J. L. Moreland, *The VRML Sourcebook*. New York: Wiley, 1996.
[2] O. Avaro, P. Chou, A. Eleftheriadis, C. Herpel, and C. Reader, "The MPEG-4 system and description languages," *Signal Process.*, vol. 9, no. 4, pp. 385–431, May 1997.
[3] "AVID effects reference guide," Avid Media Composer and Film Composer, Release 5.50, June 1995.
[4] J. R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. C. Jain, and C. Shu, "Virage image search engine: An open framework for image management," in *Proc. Symp. Electronic Imaging: Science and Technology—Storage & Retrieval for Image and Video Databases IV*, IS&T/SPIE, Feb. 1996.
[5] N. Beckmann, H. P. Kriegel, R. Schneider, and B. Seeger, "The R* tree: An efficient and robust access method for points and rectangles," in *Proc. ACM SIGMOD, Int. Conf. Management of Data*, 1990, pp. 322–331.
[6] M. Beigi, A. Benitez, and S.-F. Chang, "MetaSEEk: A content-based meta search engine for images," in *Proc. SPIE Conf. Storage and Retrieval for Image and Video Database*, San Jose, CA, Feb. 1998 (see also Columbia Univ./CTR Tech. Rep. CTR-TR #480–97–14).
[7] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ: Prentice-Hall, 1971.
[8] J. B. Black and R. McClintock, "An interpretation construction approach to constructivist design," in *Constructivist Learning Environments: Case Studies in Instructional Design*, B. G. Wilson Ed. Englewood Cliffs, NJ: Educational Technology, 1995, pp. 25–31.
[9] M. G. Brown, J. T. Foote, G. J. F. Jones, K. S. Jones, and S. J. Young, "Open-vocabulary speech indexing for voice and video mail retrieval," in *Proc. ACM Multimedia Conf.*, Boston, MA, Nov. 1996.
[10] S.-F. Chang, "Compressed-domain techniques for image/video indexing and manipulation," in *Proc. IEEE Int. Conf. Image Processing (ICIP'95)*, Washington DC, Oct. 1995.
[11] S.-F. Chang, W. Chen, H. J. Meng, H. Sundaram, and D. Zhong, "VideoQ-an automatic content-based video search system using visual cues," in *Proc. ACM Multimedia 1997*, Seattle, WA, Nov. 1997 (see demo: http://www.ctr.columbia.edu/videoq).
[12] S.-F. Chang, A. Eleftheriadis, and D. Anastassiou, "Development of Columbia's video on demand testbed," *Signal Process.*, vol. 8, no. 3, pp. 191–207, Apr. 1996.
[13] S.-F. Chang, A. Eleftheriadis, D. Anastassiou, S. Jacobs, H. Kalva, and J. Zamora, "Columbia's VoD and multimedia research testbed with heterogeneous network support," *J. Multimedia Tools Applicat.*, vol. 5, no. 2, pp. 181–184, Sept. 1997.
[14] S.-F. Chang and D. G. Messerschmitt, "Manipulation and compositing of MC-DCT compressed video," *IEEE J. Select. Areas Commun.*, vol. 3, pp. 1–11, Jan. 1995.
[15] S.-F. Chang, J. R. Smith, M. Beigi, and A. Benitez, "Visual information retrieval from large distributed on-line repositories," *Commun. ACM*, vol. 40, no. 12, pp. 63–71, Dec. 1997.
[16] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[17] D. Drelinger and A. E. Howe, "Experiences with selecting search engines using meta-search," *ACM Trans. Inform. Syst.*, 1997, to be published.
[18] "Harmonised standards for the exchange of television programme material as bit streams," Preliminary Rep., Joint EBU/SMPTE Task Force, Apr. 1997 (see also http://www.ebu.ch/pmc_es_tf.html).
[19] A. Eleftheriadis, "The MPEG-4 system description language: From practice to theory," in *Proc. 1997 IEEE Int. Conf. Circuits and Systems*, Hong Kong, June 1997.
[20] ____, "Flavor: A language for media representation," in *Proc. ACM Multimedia'97 Conf.*, Seattle, WA, Nov. 1997, pp. 1–9.
[21] Excaliber System. [Online]. Available WWW: http://www.excalib.com/rev2/products/vrw/vrw.html.
[22] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos, "Fast subsequence matching in time-series databases," in *Proc. ACM SIGMOD*, Minneapolis, MN, May 1994, pp. 419–429.
[23] Y. Fang and A. Eleftheriadis, "A syntactic framework for bitstream-level representation of audio-visual objects," in *Proc. 3rd IEEE Int. Conf. Image Processing (ICIP'96)*, Lausanne, Switzerland, Sept. 1996.
[24] Y. Fisher, Ed., *Fractal Image Compression*. New York: Springer-Verlag, 1995.
[25] Flavor. [Online]. Available WWW: http://www.ee.columbia.edu/flavor.
[26] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by image and video content: The QBIC system," *IEEE Comput. Mag.*, vol. 28, pp. 23–32, Sept. 1995.
[27] J. D. Foley, A. V. Dam, S. K. Feiner, J. F. Hughes, and R. L. Phillips, *Introduction to Computer Graphics*. Reading, MA: Addison-Wesley, 1993.
[28] D. Forsyth and M. Fleck, "Body plans," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Puerto Rico, June 1997.
[29] C. Frankel, M. J. Swain, and V. Athitsos, "Webseer: An image search engine for the world wide web," Department of Computer Science, University of Chicago, Chicago, IL, Tech. Rep. TR-96-14, July 31, 1996.
[30] J. H. Friedman, J. L. Bently, and R. A. Finkel, "An algorithm for finding best matches in logarithmic expected time," *ACM Trans. Mathematical Software*, vol. 3, no. 3, pp. 209–226, Sept. 1977.
[31] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Boston, MA: Kluwer Academic, 1992.
[32] A. S. Glassner, *Principles of Digital Image Synthesis*, vols. 1 and 2. Los Altos, CA: Morgan Kaufmann, 1995.
[33] J. Gosling, B. Joy, and G. Steele, *The Java Language Specification*. Reading, MA: Addison-Wesley, 1996.
[34] A. Gupta and R. Jain, "Visual information retrieval," *Commun. ACM*, vol. 40, no. 5, pp. 70–79, May 1997.
[35] A. Guttman, "R-trees: A dynamic index structure for spatial indexing," in *Proc. ACM SIGMOD, Int. Conf. Management of Data*, 1984, pp. 47–54.
[36] J. Hafner, H. S. Sawhney, W. Equitz, M. Flickner, and W. Niblack, "Efficient color histogram indexing for quadratic form distance functions," *IEEE Trans. Pattern Anal. Machine Intell.*, July 1995.
[37] B. G. Haskell, A. Puri, and A. N. Netravali, *Digital Video: An Introduction to MPEG-2*. New York: Chapman and Hall, 1997.
[38] A. G. Hauptmann and M. Smith, "Text, speech and vision for video segmentation: The informedia project," in *Proc. AAAI*

Fall Symp., Computational Models for Integrating Language and Vision, Boston, MA, Nov. 10–12, 1995.
[39] J. Huang, S. R. Kumar, and M. Mitra, "Combining supervised learning with color correlograms for content-based image retrieval," in *Proc. ACM Multimedia'97*.
[40] "Special Issue on MPEG-4," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 1, Feb. 1997.
[41] Institute for Learning Technologies. The Eiffel project: New York City's small schools partnership technology learning challenge. [Online]. Available WWW: http://www.ilt.columbia.edu/eiffel/eiffel.html.
[42] Internet 2 Project. [Online]. Available WWW: http://www.internet2.edu.
[43] M. Irani, H. S. Sawhney, R. Kumar, and P. Anandan, "Interactive content-based video indexing and browsing," in *Proc. IEEE Multimedia Signal Processing Workshop*, Princeton, NJ, June 1997.
[44] "Information technology—Coding of moving pictures and associated audio for digital storage media at up to about 1, 5 Mbit/s," ISO/IEC 11172 International Standard (MPEG-1), 1993.
[45] "Information technology—Generic coding of moving pictures and associated audio (also ITU-T Rec. H.262)," ISO/IEC 13818 International Standard (MPEG-2), 1995.
[46] "Virtual reality modeling language," ISO/IEC 14472 Draft International Standard, 1997.
[47] ISO/IEC JTC1/SC29/WG11 (MPEG). [Online]. Available WWW: http://www.cselt.it/ mpeg.
[48] "Context and objectives (v. 2)," ISO/IEC JTC1/SC29/WG11, MPEG-7: Sevilla, Italy, Feb. 1997.
[49] "Description of MPEG-4," ISO/IEC JTC1/SC29/WG11 N1410, Oct. 1996.
[50] "MPEG-4 requirements version 4.0," ISO/IEC JTC1/SC29/WG11 N1727, July 1997.
[51] "MPEG-4 applications," ISO/IEC JTC1/SC29/WG11 N1729, July 1997.
[52] "MPEG-4 overview," ISO/IEC JTC1/SC29/WG11 N1730, July 1997.
[53] "MPEG-4 audio working draft version 4.0," ISO/IEC JTC1/SC29/WG11 N1745, July 1997.
[54] "MPEG-4 visual working draft version 4.0," ISO/IEC JTC1/SC29/WG11 N1797, July 1997.
[55] "MPEG-4 systems working draft version 5.0," ISO/IEC JTC1/SC29/WG11 N1825, July 1997.
[56] "Video codec for audio visual services at $p \times 64$ kbit/s," International Telecommunications Union, Geneva, Switzerland, ITU-T Recommendation H.261, 1990.
[57] C. E. Jacobs, A. Finkelstein, and D. H. Salesin, "Fast multiresolution image querying," in *Proc. ACM SIGGRAPH*, Aug. 1995, pp. 277–286.
[58] R. A. Jarvis, "A perspective on range finding techniques for computer vision," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 5, pp. 122–139, Mar. 1983.
[59] N. S. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. Englewood Cliffs, NJ: Prentice-Hall, 1984.
[60] H. Kalva, S.-F. Chang, and A. Eleftheriadis, "DAVIC and interoperability experiments," *J. Multimedia Tools Applicat.*, vol. 5, no. 2, pp. 119–132, Sept. 1997.
[61] T. Kanade, "Development of a video rate stereo machine," in *Proc. ARPA Image Understanding Workshop*, Nov. 1994, pp. 549–557.
[62] K. A. Kastens, D. van Esselstyn, and R. O. McClintock, "'Where are we?' An interactive multimedia tool for helping students 'translate' from maps to reality and vice versa," *J. Geoscience Educ.*, vol. 44, pp. 529–534, 1996.
[63] K. Kozel. (Sept. 1996). The object of object-oriented authoring. *CD-ROM Professional*. [Online]. Available WWW: http://www.onlineinc.com/cdrompro/0996CP/kozel9.html.
[64] ____. (July 1997). The classes of authoring programs. *EMedia Professional*. [Online]. Available WWW: http://www.onlineinc.com/ emedia/EM- tocs/emtocjul.html.
[65] C.-S. Li, L. Bergman, S. Carty, V. Castelli, S. Hutchins, L. Knapp, I. Kontoyiannis, J. Robinson, R. Ryniker, J. Shoudt, B. Skelly, and J. Turek, submitted for publication.
[66] M. Li and P. Vitanyi, *An Introduction to Kolmogorov Complexity and Its Applications* New York: Springer-Verlag, 1993.
[67] J. Liang, "Highly scalable image coding for multimedia applications," in *Proc. ACM Multimedia Conf.*, Seattle, WA, Nov. 1997.
[68] T. D. C. Little and A. Ghafoor, "Spatio-temporal composition of distributed multimedia objects for value-added networks," *IEEE Comput. Mag.*, pp. 42–50, Oct. 1991.
[69] R. McClintock, *Power and Pedagogy: Transforming Education Through Information Technology*. New York: Institute for Learning Technologies, 1992.
[70] R. McClintock, F. A. Moretti, L. Chou, and T. de Zengotita, *Risk and Renewal: First Annual Report—1991–1992: The Phyllis and Robert Tishman Family Project in Technology and Education* New York: New Laboratory for Teaching and Learning, The Dalton School, 1992.
[71] J. Meng and S.-F. Chang, "CVEPS: A compressed video editing and parsing system," in *Proc. ACM Multimedia Conf.*, Boston, MA, Nov. 1996 (see demo: http://www.ctr.columbia.edu/webclip).
[72] J. Meng, D. Zhong, and S.-F. Chang, "A distributed system for editing and browsing compressed video over the network," in *Proc. IEEE 1st Multimedia Signal Processing Workshop*, Princeton, NJ, June 1997.
[73] ____, "WebClip: A WWW video editing/browsing system," in *Proc. IEEE 1st Multimedia Signal Processing Workshop*, Princeton, NJ, June 1997 (see demo: http://www.ctr.columbia.edu/webclip).
[74] T. P. Minka and R. Picard, "Interactive learning using a 'society of models'," MIT Media Laboratory Perceptual Computing Section Tech. Rep. 349.
[75] ____, "An image database browser that learns from user interaction," MIT Media Laboratory and Modeling Group Tech. Rep. 365, 1996.
[76] R. Mohan, "Text based search of TV news stories," in *Proc. SPIE Photonics East Int. Conf. Digital Image Storage & Archiving System*, Boston, MA, Nov. 1996.
[77] V. Nalwa, *A True Omnidirectional Viewer*. Holmdel, NJ: AT&T Bell Laboratories, 1996.
[78] S. K. Nayar, "Catadioptric omnidirectional cameras," Tech. Rep., Oct. 1996 (see demo: http://bagpipe.cs.columbia.edu/Omnicam).
[79] S. K. Nayar, M. Watanabe, and M. Noguchi, "Real-time focus range sensor," *IEEE Trans. Pattern Anal. Machine Intell.*, Dec. 1996.
[80] A. N. Netravali and B. G. Haskell, *Digital Pictures: Representation, Compression, and Standards*, 2nd ed. New York: Plenum, 1995.
[81] V. E. Ogle and M. Stonebraker, "Chabot: Retrieval from a relational database of images," *IEEE Comput. Mag.*, vol. 28, no. 9, pp. 40–48, Sept. 1995.
[82] T. A. Ohanian, *Digital Nonlinear Editing: New Approaches to Editing Film and Video*. Boston, MA: Focal, 1993.
[83] W. Pennebaker and J. Mitchell, *The JPEG Still Image Data Compression Standard*. New York: Van Nostrand, 1993.
[84] A. Pentland, R. W. Picard, and S. Sclaroff, "Photobook: Tools for content-based manipulation of image databases," in *Proc. SPIE Storage and Retrieval for Image and Video Databases II*, Bellingham, WA, 1994, vol. 2185, pp. 34–47.
[85] E. G. M. Petrakis and C. Faloutsos, "Similarity searching in medical image databases," University of Maryland Tech. Rep. UMD: CS-TR-3388, UMIACS-TR-94-134 (extended version).
[86] Y. Rui, T. Huang, S. Mehrotra, and M. Ortega, "A relevance feedback architecture for content-based multimedia information retrieval systems," in *Proc. CVPR'97 Workshop Content-Based Image and Video Library Access*, June 1997.
[87] B. Shahraray and D. C. Gibbon, "Automatic generation of pictorial transcript of video programs," in *Proc. SPIE*, vol. 2417, 1995, pp. 512–518.
[88] J. M. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. Signal Processing*, vol. 41, pp. 3445–3462, Dec. 1993.
[89] S. Shatford, "Analyzing the subject of a picture: A theoretical approach," *Library of Congress Cataloging Classification Quart.*, vol. 6, 1985.
[90] "Special issue on MPEG-4, part 1: Invited papers," *Signal Process.*, vol. 10, nos. 1–3, May 1997.
[91] "Special issue on MPEG-4, part 2: Submitted papers," *Signal Process.*, vol. 10, no. 4, July 1997.
[92] J. R. Smith and S.-F. Chang, "VisualSEEk: A fully automated content-based image query system," in *Proc. ACM Multimedia Conf.*, Boston, MA, Nov. 1996 (see demo: http://www.ctr.columbia.edu/VisualSEEk).

[93] ____, "Searching for images and videos on the world-wide web," *IEEE Multimedia Mag.*, Summer 1997.
[94] ____, "Enhancing image search engines in visual information environments," in *Proc. IEEE 1st Multimedia Signal Processing Workshop*, Princeton, NJ, June 1997.
[95] S. W. Smoliar and H. Zhang, "Content-based video indexing and retrieval," *IEEE Multimedia Mag.*, Summer 1994.
[96] D. Sow and A. Eleftheriadis, "Complexity distortion theory," in *Proc. IEEE Int. Symp. Information Theory and Its Applications*, June 1997.
[97] ____, submitted for publication.
[98] R. F. Sproull, "Refinements to nearest-neighbor searching in K-dimensional trees," *Algorithmica*, vol. 6, no. 4, pp. 579–589, 1991.
[99] R. K. Srihari, "Automatic indexing and content-based retrieval of captioned images," *IEEE Comput. Mag.*, vol. 28, no. 9, pp. 49–58, Sept. 1995.
[100] J. Swartz and B. C. Smith, "A resolution independent video language," in *Proc. ACM Multimedia Conf.*, 1995.
[101] L. Torres and M. Kunt, Eds., *Video Coding: The Second Generation Approach*. Boston, MA: Kluwer Academic, 1996.
[102] M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding* Englewood Cliffs, NJ: Prentice-Hall, 1995.
[103] S. Weibel and E. Miller, "Image description on the Internet: A summary of the CNI/OCLC Image Metadata on the Internet Workshop," Sept. 24–25, 1996, *D-Lib Mag.*, Jan. 1997.
[104] World Wide Web Consortium. Synchronized multimedia activity. [Online]. Available WWW: http://www.w3.org/Audio Video/Activity.html.
[105] B. L. Yeo and B. Liu, "Rapid scene analysis on compressed videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 5, no. 6, Dec. 1995.
[106] M. M. Yeung and B. L. Yeo, "Video content characterization and compaction for digital library applications," in *Proc. SPIE, Storage and Retrieval for Still Image and Video Databases V*, vol. SPIE 3022, Feb. 1997, pp. 45–58.
[107] T. de Zengotita, R. McClintock, L. Chou, and F. A. Moretti, *The Dalton Technology Plan: Second Annual Report—1992–1993*, vol. 1, *Developing an Educational Culture of Skill and Understanding in a Networked Multimedia Environment*. New York: New Laboratory for Teaching and Learning, 1993; and vol. 2, *Proof of Concept: Educational Innovation and the Challenge of Sustaining It*. New York: New Laboratory for Teaching and Learning, 1993.
[108] D. Zhong, H. Zhang, and S.-F. Chang, "Clustering methods for video browsing and annotation," in *Proc. SPIE Conf. Storage and Retrieval for Image and Video Database*, San Jose, CA, Feb. 1996.

Shih-Fu Chang (Member, IEEE) received the Ph.D. degree in electrical engineering and computer science from the University of California, Berkeley, in 1993.

He currently is an Associate Professor in the Department of Electrical Engineering and New Media Technology Center, Columbia University, New York. He also actively participates in Columbia's Digital Library Project. His current research interests include content-based visual query, networked video manipulation, and video communication. He is particularly interested in application of content-based video processing to image/video retrieval, network resource management, and image watermarking and authentication. His group has developed several large-scale Web-based prototypes of visual information systems, including an MPEG video-editing engine, WebClip, content-based visual search engines, VideoQ and WebSEEk, and a metasearch engine for images/video, MetaSEEk. He has received two U.S. patents (with six pending) in the area of visual search and compressed video processing. He has been actively participating in the technical activities of several international publications and conferences. He also taught several short courses on visual information retrieval.

Prof. Chang has received two Best Paper awards. He received an ONR Young Investigator Award for 1998–2001, an NSF CAREER award for 1995–1998, and an IBM UPP Program Faculty Development Award for 1995–1998.

Alexandros Eleftheriadis (Member, IEEE) was born in Athens, Greece, in 1967. He received the diploma in electrical engineering and computer science from the National Technical University of Athens, Greece, in 1990 and the M.S., M.Phil., and Ph.D. degrees in electrical engineering from Columbia University, New York, in 1992, 1994, and 1995 respectively.

Since 1995, he has been an Assistant Professor in the Department of Electrical Engineering at Columbia University, where he is leading a research team working in the areas of visual information representation and compression, video communications systems (including video on demand and Internet video), distributed multimedia systems, and the fundamentals of compression. During the summers of 1993 and 1994, he was with AT&T Bell Laboratories, Murray Hill, NJ, developing low-bit-rate model-assisted video-coding techniques for video-conferencing applications. From 1990 to 1995, he was a Graduate Research Assistant in the Department of Electrical Engineering at Columbia University. He is a member of the ANSI NCITS L3.1 Committee and the ISO/IEC JTC1/SC29/WG11 (MPEG) group that develop national and international standards for audio-visual content representation and distribution.

Dr. Eleftheriadis is a member of the Association of Computing Machinery and the Technical Chamber of Greece.

Robert McClintock (Associate Member, IEEE) received the B.A. degree from Princeton University, Princeton, NJ, in 1961 and the Ph.D. degree from Columbia University, New York, in 1968.

He is a Professor of communication, computing, and technology in education at Teachers College, Columbia University, New York, and Codirector of the Institute for Learning Technologies. His scholarship covers a wide range—an intellectual biography of the Spanish philosopher José Ortega y Gasset, essays on the history of educational thought, and works on technology and education. Since 1986, He has directed the Institute for Learning Technologies, through which he has developed projects to prototype how children and teachers will interact with advanced digital curricular resources over the information infrastructure—the Harlem Environmental Access Project, a collaboration with the Environmental Defense Fund, with TIIAP support; the Living Schoolbook Project, a collaboration with the Syracuse School of Education, with NYS Science and Technology Foundation support; Reinventing Libraries, a program to redefine the way school libraries use advanced media resources; and the Eiffel Project, a five-year U.S. Challenge Grant for Technology in Education, which will use new media to support school reform in some 80 New York City schools and community organizations.