derives from its relations with other images, which vary according to the query's particular circumstances. This implies that the semantics of images is at least in part functional and that a query process for image databases should manipulate similarity functions. An image database should include a complete algebra of similarity functions and treat similarity functions as first-class data.

Second, the semantics of the image's descriptors (features) should be specified, as much as possible, through a discourse (that is, through algebraic, logic, and functional means). However, this formalization will never be sufficient to delimit a semantics of interest—it will merely help in practical aspects of database organization[8] and support the user's true semantic-generating activity.

Finally, an image always has a meaning relative to the practices and social codes of a specific user. For example, two people in a picture can be judged too close (and therefore in a situation of intimacy) for an American viewer, but at a fair distance (and therefore in a situation of formality) for an Italian viewer, simply because the social code of spatial configurations is different in the two cases. In this sense, the goal of the interaction between the user and database is not so much to retrieve images based on a preexisting semantics but to create image semantics. The interaction itself is not configured as a query but as a navigation in which the user dictates similarities and associations between images and, through this activity, reorganizes the database to embody the desired semantic.

It is essential, for instance, that through the use of appropriate interfaces,[9] the user can decide which images are similar. This activity lets the database adapt its similarity measure to that which the user has in mind for that particular query. Consequently, the database can build, through repeated iterations, the semantics that the user has in mind for that particular query.

Relevance feedback has been a first step in this direction, but it is clear that to let alternative semantics emerge from the interaction between the user and database, the connection between the two must be much deeper. The user needs expressive means more powerful than simply selecting positive or negative examples, and the whole data organization inside the database should depend on the status of the interaction with the user.

The challenges that this organization will pose are at the boundary between database theory, image analysis, knowledge representation, and human–machine interaction. Developing solutions from such a maelstrom of different technical cultures and orientations will be an interesting and exciting experience.

References

1. F. Merrel, *Semiosis in the Post-Modern Age*, Purdue Univ. Press, Lafayette, Ind., 1995.
2. U. Eco, *A Theory of Semiotics*, Indiana Univ. Press, Bloomington, Ind., 1976.
3. D. Lenant, *The Dimensions of Context-Space*, tech. report, Cycorp, 1998.
4. A. Albano, G. Ghelli, and R. Orsini, "Fibonacci: A Programming Language for Object Databases," *The VLDB J.*, vol. 4, no. 3, July 1995, pp. 403–444.
5. P. Buneman et al., "Principles of Programming with Complex Objects and Collection Types," *Theoretical Computer Science*, vol. 149, no. 1, Sept. 1995, pp. 3–48.
6. K. Didrich et al., "Programming in the Large: The Algebraic-Functional Language Opal 2," *Implementation of Functional Languages* (IFL'97), Lecture Notes in Computer Science, vol. 1467, Springer-Verlag, Berlin, 1998, pp. 323–338.
7. S. Santini, *Exploratory Image Databases: Content-Based Retrieval*, Academic Press, San Diego, Calif., 2001
8. S. Santini and A. Gupta, "An Algebra of Wavelet Features," *IEEE Int'l Conf. Multimedia and Expo* (ICME 2001), IEEE Press, Piscataway, N.J., 2001.
9. S. Santini, A. Gupta, and R. Jain, "Emergent Semantics through Interaction in Image Databases," *IEEE Trans. Knowledge and Data Eng.*, vol. 13, no. 3, May/June 2001, pp. 337–351.

Media Information Spaces—A Semantic Challenge

Frank Nack, *CWI, Amsterdam*

The information society is leaving behind the cyberspace based on a hybrid system of traditional media (telephone, cinema, TV, theatre, museum, books, newspapers, and so forth) and digital information technology (networked and storage intensive computers, CD-ROMs, DVD, IP-telephony, Webcams, MIDI, and so forth). Rather, it is entering a knowledge space that facilitates new forms of creativity, knowledge exploration, and social relationships mediated through communication networks (including hypertext, interactive multimedia, interactive games, virtual reality, simulations, and augmented reality).

Such an interactive, open, and multimodal environment sustains the activation of the human and the artificial system's articulation powers to communicate ideas, where verbal, gestical, musical, iconic, graphic, or sculptural expressions form the basis of adaptive discourses. A basic aspect for such a space, which supports individuals but is still communal, is that information must be made accessible that is hidden in the unified structure of the single text, image, video, audio, or tactile unit. Thus, the goal is to create an environment in which media units and the relationships among them are understood as basic elements that can interrelate to produce new meanings.

To support this process of generating meaning, interpretation, and visualization, a system must know what is contained in the different media. For visual media, however, this poses a problem. Even though an image might provide a limited amount of visual information, it contains a wealth of meaning. This functionality is based on the two formal structures that can be assigned to every perceivable object in visuals: the *signifier* (which carries the meaning) and the *signified* (which is the concept or idea signified). The relation between the two elements is not a naming-process only, as the signified resembles not a thing but a concept. Secondly, the relation between the signifier and the signified is arbitrary. It is, in particular, the arbitrariness of the relationship between signifier and signified that enables the creation of higher-order sign systems and their diversity.

Thus, visual media requires more than characterizing its visual information on a perceptual level using objective measurements, such as those based on image or sound processing or pattern recognition. Creatively reusing material for individual purposes, which usually opens up questions of aesthetics and subjective interpretation, has a strong influence on the descriptions and annotations of visual media data, either created during the data's production process or added later. Providing semantic, episodic, and technical representation structures that can change and grow over time is important. This also requires adaptable relations between the different type of structures.

The Semantic Web

The Semantic Web is a first step toward addressing these problems (www.semanticweb.org). It should bring machine-processable content to Web pages, thus extending the current Web. The idea is to add ontology-based metadata to text or HTML documents to improve accessibility and provide a means for reasoning about the content. The applied technology is XML-based, which facilitates structural, cardinality, and datatyping constraints (XML Schema) on textual documents, allowing a comparison on structural levels. Richer semantic descriptions can be provided either as relation-oriented schemata (RDF, RDF Schema) or ontology-based technology (DAML+OIL). These technologies support in-depth indexing and classification of textual documents for presentation generation and navigation purposes.

To some extent, XML-based approaches also incorporate multimedia, either in the form of presentational languages such as Synchronized Multimedia Integration Language (SMIL) (integration of media style), SVG (with CSS for graphics), and XHTML (with CSS for formatted text), or transformational methods such as XSLT (document transformation) and CSS (control of style appearance).

However, the major drawback of XML-based environments is that they don't recognize visual media's dynamic nature or its variety of data representations and their mixes.

MPEG frameworks

The Moving Pictures Expert Group is a working group of the International Organization for Standardization/International Electronics Commission. MPEG is in charge of developing standards for coded representation of digital audio and video, and it leads one of the broadest efforts in the direction of complex media content modeling. It aims to provide a framework for interoperable multimedia content-delivery services.

Semantic description languages have emerged in two of its standardization activities: in MPEG-4, as the Extensible MPEG-4 Textual Format (XMT) and in MPEG-7, as the Description Definition Language (DDL)—the multimedia content description interface.

In MPEG-4, the standard for multimedia on the Web, XMT provides content authors with a textual syntax for the MPEG-4 Binary Format for Scenes (BIFS) to exchange their content with other authors, tools, or service providers. XMT is an XML-based abstraction of the object descriptor framework for BIFS animations. Moreover, it respects existing practices for authoring content, such as SMIL, HTML, or Extensible 3D by allowing the interchange of the format between a SMIL player, a Virtual Reality Modeling Language player, and an MPEG player. It does this using the relevant language representations such as XML Schema, MPEG-7 DDL, and VRML grammar. In short, XMT serves as a unifying framework for representing multimedia content where otherwise fragmented technologies are integrated and the interoperability of the textual format between them is bridged.

The MPEG-7 group's objective is to standardize ways of describing different types of multimedia information. The emphasis is on audio–visual content with the goal of extending the limited capabilities of proprietary solutions to identify content by providing a set of *description schemes* and *descriptors* to make various types of multimedia content accessible. In this context, a description scheme specifies the structure and semantics of the relationships between its components, which might be both descriptors and description schemata. A descriptor defines the syntax and the semantics of a distinctive characteristic of the media unit to be described, such as an image's color, a speech segment's pitch, an audio segment's rhythm, a video's camera motion or style, a movie's actors, and so forth. Descriptors and description schemata are represented in the MPEG-7 DDL. The current version of the DDL is based on XML Schema, which provide a means of describing temporal and spatial features of audio–visual media as well as connecting these spatio-temporal descriptions within the media. The DDL also provides the necessary mechanisms for extending and refining existing description schemata and descriptors and to define new schemata or descriptors if required.

Current problems

Problems exist with using MPEG-7 as the basis for a dynamic media-based knowledge space. First, MPEG-7 is hierarchy centered. This means that a description of data in MPEG-7 is understood as one document that applies a tree structure. The schemata for this document type are fixed and cannot be altered. This linear approach is not astonishing, because efficient access and retrieval was and still is the driving development force of the MPEG-7 standardization effort. However, this approach is far too restrictive; any form of annotation is necessarily imperfect, incomplete, and preliminary, because annotations accompany and document the progress of interpreting and understanding a concept. Graphs, which form the basis of semantic networks, provide better support for carrying out this incomplete task over time.

Related to this problem is the conceptual idea in MPEG-7 of two general description types: complete descriptions (which use the MPEG-7Main as the root element) and partial description units (which use the MPEG-7Unit as the root element). Distinguishing between a complete and fragmental description is purely academic and adds an unnecessary level of complexity.

Another problem is the great number of MPEG-7 schemata—not so much because of their number, which is unavoidable, but because of their interlocked nature, which makes using schemata in isolation difficult.

Finally, it has also become increasingly clear that we need a machine-understandable representation of the semantics associated with MPEG-7 description schemes and descriptors. This representation would enable the interoperability and integration of MPEG-7 with metadata descriptions from other domains. MPEG-7 is currently developing description schemata mainly for the film and broadcasting domain, and to accomplish this, MPEG-7 requires a common understanding of the semantic relationships between metadata terms from different domains. XML Schema, and hence MPEG-7's DDL, provide little support for expressing semantic knowledge, but RDF Schema might. Jane Hunter and Carl Lagoze offer an example for interoperability between application profiles in RDF and XML Schema.[1]

Striving to be a highly interoperable standard among well-known industry standards and other related standards of different domains is a courageous and farsighted step for a group mainly known for its concern with efficient audio–visual coding at the bit level. Moreover, the textual representations in

MPEG-4 and MPEG-7 not only support the current trend in content description toward XML as the accepted standard, but they also point to new ground. Because textual representations allow a symbolic representation of multimedia content by expressing relations between elements—synchronized with the different modalities of multimedia data—it is now possible to model central aspects of how humans try to make sense of complex systems.

So, has the paradigm change in multimedia computing happened yet? Not really, but we're moving in the right direction. The real challenges are still ahead of us—generating and using quality metadata.

It took nearly 30 years of steady infiltration of technological advances in everyday production environments—such as nonlinear video-editing systems, image-editing tools, audio systems, and Web presentation technology—to communicate ideas in forms other than text. And still, the technology follows the strains of traditional written communication by supporting the linear representation of an argument, which results in a final multimedia product of context-restricted content. Thus, we face the paradoxical situation that although there are more possibilities than ever to assist in the creative development and production processes of media, we still lack adaptive environments that can serve as an integrated information space for use in distributed productions, research, restructuring (such as by software agents), or direct access and navigation.

We need systems for authoring media that let people use their creativity in familiar ways and their human actions to extract the significant syntactic, semantic, and semiotic aspects of the media's content to construct descriptions based on a formal language. There is much evidence that manual labor can provide a great deal of useful annotation.[2–4] We also need systems that manage independent media objects and representations for use in many different productions with a potentially wide range of applications.

Yet, if we only had the information gathered during the production of media, including its reuse and modifications, we would still lack knowledge about the material's potential intrinsic meanings. Thus, it is important to make people aware that the notion of a completed work vanishes in such a system and leaves space for a creative and productive cycle, a living environment allowing all sorts of processes. These spaces are for investigation based on an interpreting, associative method rooted in a discourse-oriented collective interpretation of questions that, by following the branches of interdependencies, compare the most diverse theories.

References

1. J. Hunter and C. Lagoze, "Combining RDF and XML Schemas to Enhance Interoperability Between Metadata Application Profiles," *Proc. 10th Int'l WWW Conf.*, 2001, pp. 456–466.

2. C. Dorai and S. Venkatesh, "Bridging the Semantic Gap in Content Management Systems: Computational Media Aesthetics," *Proc. 1st Conf. Computational Semiotics for Games and New Media* (COSIGN 2001), 2001, pp. 94–99; www.kinonet.com/conferences/cosign2001/program.html (current Jan. 2002).

3. A.T.G. Schreiber et al., "Ontology-Based Photo Annotation," *IEEE Intelligent Systems*, vol. 16, no. 3, May/June 2001, pp. 66–74; http://computer.org/intelligent/ex2001/x3066abs.htm (current Jan. 2002).

4. F. Nack and W. Putz, "Designing Annotation Before It's Needed," *Proc. 9th ACM Multimedia Conf.*, ACM Press, New York, 2001, pp. 251–260; http://acm.org/sigs/sigmm/MM2001/ep/toc.html#Wp1 (current Jan. 2002).

Emergent Semantics

Luc Steels, *University of Brussels AI Lab and Sony Computer Science Lab, Paris*

Every computer scientist knows that we can only process information when the information is somehow represented—there's no computation without representation. Traditionally, human programmers have designed the representations. They select what aspects of the domain are relevant and thus must be made explicit, and they design appropriate data structures that efficiently support the processing required for a task. This works reasonably well, but we need a massive amount of programs these days, making it difficult to keep up. Moreover, users want their programs to adapt to new tasks and a changing world. This raises the question of whether computer systems can develop and adapt representations.

A typical example is Web applications, which must cope with constantly changing information sources (material appears and disappears without any central control) and needs (the Web touches on all aspects of human life and is therefore basically open-ended). Another example is autonomous robots, which must operate in an open-ended and unpredictable world in which new tasks can arise that the designers could not have foreseen.

The origin of representation has been a central topic in AI research from the beginning—it is a problem that human biology has had to solve as well. The question is usually studied under the heading of machine learning and is far from resolved. Indeed, there is a profound paradox.

Computation requires a representation, but how can this computation generate its own representation? A representation casts a frame on the world, but this frame is a strength as well as a limitation. Stepping out of the frame is like jumping out of a hoolahoop while holding it. As Ludwig Wittgenstein put it, "The limits of my language mean the limits of my world."

We can schematically classify efforts to understand the origins of representations into two approaches: induction and selection. I propose a third alternative, which relies on interaction, construction, and communication.

Induction

The inductive approach is the best known and furthest developed, having been explored in the fields of statistical-pattern recognition,[1] symbolic machine learning,[2] and neural-network research.[3] A large training set must be available, and the inductive process goes over these data to find what is essential and what is contingent. Either the process is supervised, in the sense that it receives feedback about what it needs to learn, or it is unsupervised, in which case it attempts to detect the natural classes or regularities in the data. In the past decade, researchers have developed a wealth of induction algorithms, and many applications have been demonstrated for more compact coding of the data, finding similarities, learning inference rules, data mining, and so forth.

However, some fundamental limitations have come up as well, in the sense that the intervention of a human designer is much greater than hoped for. The designer must assemble an adequate set of training data, which she must prepare carefully. Often she must choose the outline of the repre-