INTELLIGENT RETRIEVAL
Open Journal

Review

New Frontiers for Intelligent Content-Based Retrieval

Ana B. Benitez[1*]; John R. Smith[2]

[1]*Department of Electrical Engineering, Columbia University, New York, NY 10027, USA*
[2]*IBM T. J. Watson Research Center, New York, NY 10532, USA*

[*]***Corresponding author***
Ana B. Benitez
Department of Electrical Engineering, Columbia University, New York, NY 10027, USA

Article information
Received: July 20th, 2023; **Revised:** August 16th, 2023; **Accepted:** September 21st, 2023; **Published:** October 12th, 2023

Cite this article
Benitez AB, Smith JR. New frontiers for intelligent content-based retrieval. 2023; 1(1). doi: https://doi.org/10.70705/ppp.ir.2023.v01.i01.pp42-49

ABSTRACT

New developments in intelligent infrastructure-based content retrieval systems are explored in this article. In this context, "intelligence" is defined as a system's capacity to reason and learn effectively, use context to its advantage, construct and maintain models of the world or specific situations, and make use of dynamic representations of information. In order to build efficient systems for retrieving audiovisual material at semantic levels comparable to human perception and cognition, we contend that these components are necessary. We survey studies from cognitive psychology, AI, semiotics, and computer vision that provide light on the nature of intelligence and how intelligent systems are built. We go on to talk about how content-based retrieval systems might take use of new possibilities made possible by some of the key concepts in these areas. Lastly, we highlight a few of our initiatives in these areas. A multimedia knowledge presentation system called MediaNet and some MPEG-7 description tools that allow intelligent content-based retrieval are specifically introduced.

Keywords
Intelligent content-based retrieval; Cognitive psychology; Artificial intelligence; Semiotics; Dynamic knowledge; Context; Reasoning; Learning; MediaNet; MPEG-7.

INTRODUCTION

The first generation of multimedia information systems, made possible by recent studies on audiovisual content analysis, can already handle content-based retrieval, automated but limited scene and object categorization, and basic learning mechanisms including user relevance feedback. We still haven't seen these features significantly distinguish operational multimedia information systems, but they're a big step forward from just using textual keywords for indexing and retrieval. Developing these technologies has been a good start, but we still have a long way to go.

Full integration of intelligence into systems will characterize the future generation of multimedia information systems, in our opinion. Research in areas like computer vision, cognitive psychology, AI, and semiotics will propel this advancement by building upon the groundwork laid by signal processing and pattern analysis. In our opinion, for content-based retrieval systems of the future to be useful, they must have the ability to grasp audiovisual information at a higher semantic level and communicate with the user. Object and scene identification, complicated scene interpretation, sophisticated reasoning and learning, and many other well-known challenges are naturally brought to light by this.

2. RECENT WORK ON CONTENT-BASED RETRIEVAL

Visage 2, QBIC 11, and VisualSeek 31 are just a few examples of the many content-based retrieval (CBR) systems that have investigated the prospect of indexing visual material such as images and videos using basic visual attributes. First, these systems automatically extract characteristics from the visual data. Second, they index the extracted descriptors for rapid access. Third, they get the visual data by querying and matching the descriptors. There has been an attempt to go beyond these fundamental capabilities and provide relevant feedback in order to improve searches and understand the user's possible search terms via examples 26.

Recent years have seen concentrated efforts to automate the production of certain semantic labels that may make substantial contributions to visual data retrieval. To address the who, what, when, and where of visual material, there has been a lot of recent work on topics like portrait vs. landscape detection, indoor vs. outdoor classification, city vs. landscape classification, sunset vs. woodland classification, and many more. Traditional machine learning algorithms are the backbone of most of these approaches, and they have shown promise on limited and sometimes biased test sets. These are great first steps, but they won't get you very far towards mastery of

the material.

3. UNDERSTANDING AND CONSTRUCTION OF INTELLIGENCE

Intelligence in humans is defined as the mental capabilities to perceive, learn, remember, behave, and reason. These innate skills are used by humans in everyday life. By examining insights on human intelligence, such as those provided by psychology, we hope to better understand users of content-based retrieval systems (human intelligence) and construct more intelligent content-based retrieval systems (system intelligence). We hope to gain additional insight by studying the important principals and developments in artificial intelligence, semiotics, and computer vision.

5.1. Psychology

The goal of psychology is to understand human intelligence. Two important trends can be distinguished in psychology: behaviorism and cognitive psychology. Behaviorism investigates the correlation between percepts (information acquired from senses) and the resulting responses and human actions rejecting any theory involving specific mental processes to describe human behavior. On the other hand, cognitive psychology models the brain as an information processing system. The field of cognitive psychology considers human behavior to be a result of mental processes such as beliefs, goals, and reasoning. Developments in cognitive psychology have been a dominant influence on the foundations of artificial intelligence and semiotics.

Cognitive psychology has provided ample support for the notion that human knowledge consists of not only nodes of a textual nature but also nodes of audio-visual nature. The cognitive model proposed in 14 uses text and images to represent information about objects. Johnson-Laird 12 asserts that mental representations include text, static images, and dynamic visual and auditory content. Rumerlhart et al. 27 states that aspects of the world may be represented through multiple representational formats taking advantages of the strengths of each representation system. Other psychological studies mentioned in 8 have provided the following important insights into human intelligence: (1) humans have multiple models of the world that may be sometimes incoherence and indexed by modality; (2) humans have a distributed control center; and (3) humans are not good at performing all the tasks.

5.2. Artificial Intelligence

To comprehend intelligent beings and to build intelligent systems are the principal goals of the area known as artificial intelligence (AI). There are two primary schools of thinking when it comes to AI: those that emphasize action over cognition and those that emphasize human performance over reason, or the ideal definition of intelligence. The difference between symbolic and non-symbolic or intermediate techniques is another key point (6, 7).

Any physical symbol system possesses the required and enough means for general intelligent behavior, according to the physical symbol system hypothesis, which is followed by symbolic methods. Symbols and the processes that govern them are the building blocks of a physical symbol system. A symbol is an abstract representation of a real-world experience or activity. The term "car" stands for the idea of a tangible vehicle, for instance. Physical symbol systems are the backbone of AI methods for rational action and thought. To simulate human behavior and thought, however, researchers have looked at symbolic, non-symbolic, and, more lately, intermediate techniques.

The creation of knowledge representation models based on computers and the computerization of reasoning and logic systems are two major contributions of symbolic approaches to AI. The development of increasingly sophisticated reactive systems based on processed or direct perceptions has been spurred by non-symbolic and intermediate methods. A long-standing aim has been to imitate human cognitive processes in intelligent systems, such as robots, so that they can autonomously accomplish predetermined objectives by navigating and reasoning their surroundings. The next parts will go over logics, models of knowledge representation, and reactive systems.

3.2.1. Logics

By offering a vocabulary for representing world assertions and a set of rules for deducing new statements from old ones, logics attempt to mimic the laws of mind. The syntax and semantics of a representation language describe it, with the former outlining the language's structure and the latter its assertions' meaning. There are a variety of logics, each of which makes a unique claim about the universe and its contents (e.g., facts) and assertions (e.g., true/false/unknown). First-Order Predicate Logic (FOPL), more commonly known as First-Order Logic (FOL), is the most popular and well-understood branch of logic. It presupposes two things: first, that facts, objects, and relations among them exist; and second, that there are truth, false, and unknown beliefs. The world's propositional assertions, whether vocal or written, are represented by logics.

3.2.2. Knowledge Representation Models

Evidence suggests that scripts, frames, and semantic networks are viable options for representing knowledge. Using these techniques, any logic, like FOL, might be used to define knowledge bases. To illustrate the idea that "Bill is a person," a semantic network 25 may look like this: "Bill Clinton Node - Is-A Arc - Person Node." Here, "node" stands for an item, concept, or circumstance, and "arc" for a link between nodes. Semantic networks aren't very expressive, lacking features like negation and disjunction, despite their simplicity and support for modular inheritance.

The difference between instance and stereotype circumstances is highlighted by frames and scripts. In a frame 20, the nodes and relations are organized in a network with higher-level representations of situational features and lower-level information about particular oc-

currences of the scenario. You have the option to set and remove default settings for frame properties. An expanded version of a frame, a screenplay details not only the intended events and their order but also the objectives and strategies of the players. One argument against scripts and frames is that definitions are crucial. Another is that default values and cancellation are problematic. Lastly, it's hard to think about every critical detail of everyday circumstances.

The 9th Cyc Knowledge Server is an encyclopedia-style representation and reasoning system for common knowledge. With the addition of support for equality and default reasoning, CycL, version 10 of the Cyc representation language, extends FOL. Grouping and positioning facts about the world in distinct contexts allows for fast input and access to the information. Absolute time (like on January 1, 2000), type of time (like at night), absolute location (like in New York City), type of location (like outdoors), culture (like Catholicism), and subject (like about space) are 12 of the almost independent dimensions along which contexts are established. By feeding the written captions into Cyc's natural-language processor, we were able to improve picture retrieval.

Some other types of knowledge representation frameworks include visual pattern libraries and the Multimedia Thesaurus, which are based on multimedia, and text-based frameworks like WordNet. An computerized lexicon system, WordNet 19 arranges English words into groups of synonyms, with each word standing in for a lexicalized idea, and connects these groups via semantic linkages. Synonymy, antonymy, hypernymy/hyponymy, meronymy/holonymy, entailment, and troponymy are the semantic links that WordNet incorporates. Table 1 provides definitions and instances of each. To supplement keyword searches and database material, WordNet has been used in the retrieval of text documents and pictures with accompanying text comments 1.

There is a web of ideas, connections between ideas, and media depictions of ideas in the Multimedia Thesaurus 16, 35. Metadata such as feature vectors and excerpts from audiovisual files serve as abstractions of real-world objects with semantic significance. Common thesaurus associations such as "related," "equivalent," and "specialization/specialization" link the ideas together. Ideas that stand in for visual patterns are created using perceptual information like color and texture elements in visual pattern libraries like the SaFe system 32 and the texture image thesaurus 17. Improved content-based search, browsing, and navigation have been achieved via the use of the Visual Pattern Libraries and the Multimedia Thesaurus (16, 17, 32).

3.2.3. Reactive Systems

A sixth kind of AI that does not depend on symbols relies only on sensory input. In reaction to data gathered by the senses, these systems execute elaborate procedures that aren't always easy to understand or defend. Despite their effectiveness in certain areas, like navigation, these systems struggle with knowledge transfer to other systems due to a lack of symbols or similar abstractions.

The method outlined in 7 is an intermediary step toward instantiating symbolic representations from sensory input via recursive generalization processes (i.e., grouping according to similarity). Integrating supplementary sensory and motor skills, experiencing a body and physical coupling, social interaction with other systems and humans for learning and perfecting skills, and the progressive development of system skills are the four essences of intelligence proposed in this work.

5.3. Semiotics

In semiotics, signs and sign systems are analyzed with the purpose of assigning specific meanings, such as the word "car" and the idea of a physical entity called a car. Artificial logics, musical languages, and conversational languages are all examples of sign systems. Situational analysis offers a more suitable definition of semiotics for our purposes: "Theoretical frameworks for the study and improvement of formal tools for the acquisition, representation, organization, generation, communication, and use of knowledge are the focus of semiotics." 18. There is a strong correlation between the second definition and the well-known division of semiotics into syntax (the "car"), semantics (the "interpretant"), and pragmatics (the "object" or "real object" car) as well as the Six-Box Diagram (Figure 1) used to model intelligent behavior and thinking.

The knowledge cycle and semiotic components are shown in Figure 1. Sensors store information about the world and its events in a symbolic form. The function of perception is to provide a structured representation of sensory data via signals (syntax). This data becomes knowledge once it is further organized and generalized. In order to make decisions, it is required to interpret the information, which is achieved by creating an interpretant by adding semantics to syntax. Actuation relies on the interpretant in the same way as knowledge generation does. The world undergoes physical and/or conceptual modifications as a result of the new knowledge's arrival. Finally, the cycle is closed when new items appear that sensors can encode.

A model of intelligence as a unit that repeatedly applies three cognitive processes—focusing attention, combinatorial search, and grouping (or generalization)—gives birth to the concept of multi-resolution is semiotics. A selection of the available data is first considered. Next, many permutations of this data are produced according to predetermined criteria of resemblance. In order to create data at the next level, the optimal combination or grouping is used. The multi-resolution framework is consistent with Gödel's incompleteness theorem, which states that certain propositions cannot be proved inside a given language and so need an additional body of knowledge, such as a meta-language, to comprehend them. A multi-resolution hierarchy of languages is the outcome of this theorem. Recent years have seen efforts to bridge the gap between multimedia information systems and semiotics. The perspective of writing and interpreting multimedia signs is used to explain some of the consequences of multimedia search in Smoliar et al. 33. Media files, including pictures and text, are thought to represent ideas about real-world things (for instance, a carrot picture and the word "carrot" represent the idea of a carrot), as well as search, which is essential for reading and writing. The semiotics paradigm put out by Joyce et al. 13 explicitly adds a second representation level to 33, allowing

for the integration of high-level information (such as "carrot") and low-level metadata (such as the color histogram generated from a picture of a carrot). At this stage components that serve as indicators of the multimedia content itself are characteristics retrieved from it (refer to Figure 2). There is high-level metadata that identifies textual characteristics and low-level information that identifies non-textual aspects. Using the Multimedia Thesaurus (16, 35) and neural-network classification agents (13), we can construct a connection between the two. Del Bimbo 5 uses the concept of narrative and discourse levels of meaning production in semiotics to automatically annotate and retrieve commercial videos. The narrative level covers the fundamental signals and their combinations, while the discourse level delves into the process of constructing stories using these aspects.

Computer vision is the construction of explicit, meaningful descriptions of physical objects in images 3. The focus of computer vision is the understanding of images rather than the processing of images and, therefore, it is concerned with both low-level features and high-level semantics of images. Computer vision includes techniques for image processing, statistical pattern recognition, geometric modeling, cognitive psychology, and artificial intelligence.

The representations of the world provided by computer vision can be categorized into iconic representations, segmented images, geometric models, and relational models 3. Iconic representations are image-like representations of the world captured by different devices and techniques. Segmented images are groupings of image regions associated with meaningful objects; the image regions are usually homogeneous with respect to some criteria (e.g., texture and motion). Geometric models capture the shape of physical objects in 2D or 3D. Relational models use knowledge representation techniques such as semantic networks to represent knowledge of the world removed from perception.

4. IDEAS FOR INTELLIGENT CONTENT-BASED RETRIEVAL

Content-based retrieval systems will not be fully effective and useful until they have the intelligence to communicate with humans in conversational languages, understand audio-visual content, and reason and plan at human levels. Natural language processors such as the one in Cyc already exist that can transform text sentences to logical statements. AI logics and knowledge representation models have been proven effective to encode knowledge and enable reasoning and planning. The missing link to enable intelligent content-based retrieval is achieving human-level understanding of audio-visual content.

AI, semiotics, and computer vision approaches give us insights on how to bridge the gap between the analysis and the understanding of the audio-visual content: first, processing the content; then building and maintaining models of the world depicted in the content; finally, interpreting the content based on the models and prior available knowledge. We think the missing piece in the puzzle is a more suited representation of the perception and the knowledge of the world that contains audio-visual content.

Cognitive psychology hints at building a society of competing and cooperating models that represent the world with nodes of textual and audio-visual nature and at treating each media different (e.g., image, text, and audio). AI approaches building on both symbolic and the non-symbolic (perceptual) are the most attractive because they try to connect what is perceived (e.g. sound) to what it is interpreted (e.g. dog barking). These approaches point at the importance of the progressive development of the systems skills through interaction with humans and other systems.

From the field of semiotics, we have learned the need for a multi-resolution representation framework and the fundamental unit to generate one level from the previous one through generation, attention focus, and combinational search. A representation language will be needed to describe the corresponding view of the world at each level. Lower levels will provide more perceptual views of the world while higher levels will progressively provide more symbolic views of the world in the AI sense. Low-level features from content-based retrieval, and iconic, segmented, and geometric representations from computer vision could work at lower levels of this representation framework while conventional AI logics and knowledge representation models would correspond to intermediate levels. However, we see the development of representation languages that capture knowledge at more perceptual and more semantic levels as an important step towards developing intelligent content-based retrieval systems.

The computation complexity of systems using AI logics and knowledge representation models quickly increases. We also envision the description of knowledge in contexts at each level for efficient use and generation of knowledge. Such an approach will also satisfy a general requirement of intelligent systems: that knowledge should be encoded in such a way that it can be transferred to other systems.

5. OUR EFFORTS TOWARDS INTELLIGENT CONTENT-BASED RETRIEVAL

Audio-visual content is typically formed from the projection of real world entities through an acquisition process involving cameras and other recording devices. In this regard, audio-visual content acquisition is comparable to the capturing of the real world by human senses. This provides a direct correspondence of human audio and visual perception with the audio-visual content 30. On the other hand, text or words in a language can be thought of as symbols for the real world entities. As a result of these and the observations in the previous section, in order to deal effectively with audio-visual material, it is necessary to model real world objects and their relationships at both the symbolic and perceptual levels.

Our efforts towards intelligent content-based retrieval has focused on two fronts: MediaNet and MPEG-7. MediaNet 4 is a multimedia knowledge representation framework addressing the problem of representing real world objects using semantic and perceptual features. The MPEG-7 standard 23 aims at standardizing tools for

describing the content of multimedia content

including the structure, the semantics, and models of the multimedia content in order to facilitate a large number of multimedia searching and filtering applications. In the following sections, we present this work and discuss how it impacts intelligent content-based retrieval.

5.1. MediaNet

MediaNet is a knowledge representation framework that uses multimedia content for representing semantic and perceptual information about the world. The main components of MediaNet include conceptual entities, which correspond to world entities, and relationships among concepts. MediaNet allows the concepts and relationships to be defined or exemplified by multimedia content such as images, video, audio, graphics, text, and audio-visual features. In designing the MediaNet framework, we have built on the basic principles of semiotics and semantic networks described in previous sections.

By integrating both conceptual and perceptual representations of knowledge, MediaNet has potential to impact a broad range of applications that deal with multimedia content at the semantic and perceptual levels such as querying, browsing, summarizing, and synthesizing multimedia. In particular, we have found that MediaNet can improve the performance of multimedia retrieval applications by using query expansion and translation across multiple content modalities.

5.1.1. MediaNet: The Multimedia Knowledge Representation Framework

MediaNet represents the world using concepts and relationships between the concepts that are defined and exemplified by multimedia content such as text, images, video sequences, and audio-visual features (signs, in semiotic terminology). In MediaNet, concepts can represent either semantically meaningful objects or perceptual patterns in the world. MediaNet models the traditional semantic relationship types such as generalization and aggregation (both semiotic principles) but adds additional functionality by modeling perceptual relationships based on feature similarity and constraints. Weights and probabilities could be assigned to concepts, relationships, and media representations in MediaNet to capture dynamic knowledge and the learning process.

An example of MediaNet is shown in Figure 3. In Figure 3, the concept Human is represented by the word "human", the image of a human, and the sound recording of a human talking; the concept Hominid is represented by the text definition "a primate of the family Hominidae" and a shape descriptor; the concept Human and the concept Hominid are relate by a semantic relationship, Specialization, and a perceptual relationship, Similar Shape, with an associated feature similarity representation.
The MediaNet framework offers functionality similar to that of a dictionary or encyclopedia and a thesaurus by defining, describing, and illustrating concepts, but also by denoting the similarity of concepts at the semantic and perceptual levels. In addition, MediaNet aims at capturing the process of producing semantics (high-level representations) from perceptual patterns (low-level representations) such as a specific color and texture pattern representing a semantic concept with a given probability.

5.1.2. Implementation of MediaNet in CBR System

By integrating both semantic and perceptual representations of knowledge, MediaNet has potential to impact a broad range of applications that deal with multimedia content at the semantic and feature levels such as multimedia query, browsing, summarization, and synthesis. An intelligent content-based retrieval system for images has been implemented by extending a typical content-based retrieval system with a MediaNet knowledge base and a query processor that translates and expands queries across multiple content modalities (see Figure 4). It is important to note that the underlying search engine in still a content-based search engine.

User

The MediaNet knowledge base was constructed semi-automatically using text annotations available for some images, the electronic lexical system WordNet, and visual feature extraction tools. First, stop words were removed from the text annotations. Then, the words in the text annotations were inputted to WordNet to obtain a list of relevant concepts and semantic relationships between them with human supervision. In this step, the senses returned by WordNet for each word were filtered by a human supervisor, who removed the ones that did not apply to the image content. For a picture of a "rock, stone", the human supervisor removed the senses "rock candy, rock", "rock music, rock", and "cradle, rock", among others. A concept was created for each remaining sense. Anotonymy, hypernymy/hyponymy, and meronymy/holonomy were the only semantic relationships used from WordNet. Finally, automatic visual feature extraction tools were used to extract features from the images. A concept was also associated the centroids of the feature descriptors of the images representing the concept.

In the current implementation, the query processor uses the MediaNet knowledge base basically to pre-process incoming queries from users. First, the query processor classifies each incoming query into a set of relevant concepts based on the media representations of the concepts (centroids and visual features of images). The initial set of relevant concepts is then extended with other semantically similar concepts. A content-based query is issued to the CB search engine for the initial user query and for each relevant concept. The feature centroids of the concept are used as the CB query for the concept. Finally, the results of all the queries are merged into a unique list for the user by taking the weighted minimum distance scores for each result image. The weights are determined based on how similar the media representations of the concepts that generated those results were to the initial user query. The query processor could also use the MediaNet knowledge base to further process the results of CB queries.

5.1.3. Evaluation of MediaNet in CBR System

We set up several experiments to evaluate MediaNet in searching for images. In particular, we compared the performance of the intelligent content-based retrieval system with the typical content-based retrieval system in Figure 4. For the intelligent content-based retrieval system, we distinguished two cases: image and text queries. The retrieval effectiveness was measured in terms of precision and recall 29. Recall and precision are standard measures used to evaluate the effectiveness of a retrieval engine. Recall is defined as the percentage of relevant images that are retrieved. Precision is defined as the percentage of retrieved images that are relevant.

The image collection selected for the experiment was 5466 images used in MPEG-7 to evaluate and compare color description technology. This collection includes photographs and frames selected from video sequences from a wide range of domains: sports, news, home photographs, documentaries, and cartoons, among others. The ground truth for 50 color queries was also generated by MPEG-7. The ground truth of each query represents a semantic, visual class and is annotated by a short textual description (e.g., "Flower Garden" and "News Anchor"). Initially, we used the color ground truth generated by MPEG-7 to compare the retrieval effectives of both systems but found it to be not suited because of its very limited semantics. We, then, generated the ground truth with relevance scores for the semantic query "tapirs". Relevance scores were assigned to images in the ground truth as follows: "1" for images of tapirs, "0.75" for images of mammals, "0.5" for images of earth animals; "0.25" for images of water and air animals; and "0" for the rest of the images.

We used the textual descriptions associated with the ground truth of the queries to construct the MediaNet knowledge base as described in the previous section. The total number of concepts derived from these textual annotations was 96. 50 of these concepts were related to other concepts by generalization/specialization relationships (hypernymy/hyponymy); 34 concepts were related to other concepts by membership, composition, or substance relationships (meronymy/hyponymy). There was only one case of antonymy. Half of the images in the ground truth were used to generate the image and feature representations of the concepts in the MediaNet knowledge base; these images were not included in the feature database of the CB search engine.

Figure 5 shows the average precision and recall for the typical and the intelligent content-based retrieval systems for the 50 MPEG-7 color queries and the semantic query "tapirs" using color histogram. For image queries, the performance of both systems is comparable for the 50 MPEG-7 queries; however, the intelligent content-based system shows a considerable improvement of retrieval effectiveness for the semantic query "tapirs". As expected, the retrieval effectiveness for text queries in the intelligent content-based retrieval engine is much lower than for image queries due to the small number of words in the MediaNet knowledge base. The results using color histogram, color coherence, wavelet texture, and Tamura texture were very similar. Although these results are very encouraging, additional experiments are needed to further demonstrate the performance gain of using MediaNet in a content-based retrieval system.

5.1.4. Future Work

Some of the future work items for MediaNet are to introduce knowledge contexts taking Cyc's contexts 15 as the starting point; to add learning, inference, and reasoning capabilities to the framework; and to further continue the evaluation of MediaNet in content-based retrieval systems and other multimedia information systems.

5.2. MPEG-7

The MPEG-7 standard 23 aims at standardizing tools for describing the content of multimedia material in order to facilitate a large number of multimedia searching and filtering applications. MPEG-7 description tools describe different aspects of multimedia material such as the features, structure, semantics, and models of multimedia content 21, 22. This section describes some the semantic and model description tools that have the highest potential to impact intelligent content-based retrieval systems.

The semantic description tools allow to represent narrative worlds depicted in or related to multimedia content in terms of semantic entities and relationships between semantic entities. Semantic entities can be objects existing in the world; events taking place in the world; abstractions, interpretations, and attributes of objects and events; and semantic times and places. A graphical example of a semantic description of a piece of audio-visual content is shown in Figure 6. This description states that Tom is playing the piano on Saturday night at Carnegie Hall. This event is interpreted as being harmonic and a tribute to Tom's mentor.

Figure 6: Semantic description of the narrative world depicted in piece of audio-visual content. The description states that Tom is playing the piano on Saturday night at Carnegie Hall. This event is interpreted as being harmonic and a tribute to Tom's mentor.

The description of semantic entities can point to the media where the semantic entities appear and contain audio-visual features of the media appearances of the semantic entities. Semantic entities can also be described by models related to audio- visual content. The MPEG-7 model description tools provide parameterized descriptions of collections or classes of audio- visual content. The models can be expressed in terms of statistics or probabilities associated with the attributes of collections of audio-visual content, or can be expressed through examples or exemplars of the audio-visual content classes.

The descriptive power and functionality of the MPEG-7 semantic and model description tools has been proven through continuous experimentation. They provide a rich framework to describe the world captured by or related to multimedia material. MediaNet knowledge bases could be encoded using these descriptions tools, which would greatly benefit the exchange and re-use of knowledge

and intelligence among multimedia applications. The components of MediaNet could be mapped to MPEG-7 semantic and model description tools as follows. Each concept in MediaNet could be a semantic entity. The text representations of a concept could be text definitions of the semantic entity. Other media representations of concepts could be described by probability models (e.g., centroids of concepts) and audio-visual examples of semantic entities. Relationships among concepts in MediaNet could be encoded as relationships among the corresponding semantic entities.

6. SUMMARY

In order to facilitate intelligent content-based retrieval systems, this work has reviewed studies in computer vision, psychology, artificial intelligence, semiotics, and the study of intelligent beings. Intelligence is defined as a system's capacity to reason and learn effectively, use context to their advantage, construct and maintain models of the world or specific situations, and make use of dynamic representations of information. Intelligent content-based retrieval systems have been covered in terms of their potential benefits and drawbacks. In this article, we also describe some of our work in these areas, including MediaNet, a framework for presenting multimedia information, and MPEG-7 description tools, which help with intelligent content-based retrieval.

REFERENCES

1. "Using Semantic Contents and WordNet in Image Retrieval" by Y. A. Aslandogan, C. Their, C. T. Yu, and N. Rishe was published in 1997 in the Proceedings of the 20th It.l ACM SIGIR Conference on Research and Development in Information Retrieval, on pages 286-295.

"Virage Image Search Engine: An Open Framework for Image Management" was published in 1996 by J. R. Bach et al. and was presented at the IS&T/SPIE conference in San Jose, California.

3. In 1982, D. H. Ballard and C. M. Brown published "Computer Vision" in the Englewood Cliffs, NJ, publication of PRENTICE-HALL INC.

The paper "MediaNet: A Multimedia Information Network for Knowledge Representation" was presented at the IS&T/SPIE-2000 conference in Boston, Massachusetts in November 2000 by A. B. Benitez, J. R. Smith, and S.-F. Chang.

6. "Expressive Semantics for Automatic Annotation and Retrieval of Video Streams" by A. Del Bimbo was presented at the ICME-2000 conference in July 2000 in New York, NY.

R. A. Brooks's "Intelligence Without Reason" was published in April 1991 in MIT AI Lab Memo 1293.

7. In "Building Brains for Bodies," published in November 1994 in Autonomous Robots, R. A. Brooks and L. A. Stein discuss the topic.

In AAAI-98, the paper "Alternate Essences of Intelligence" was presented by R. A. Brooks, C. Breazeal, R. Irie, C. Kemp, M. Marjanovic, B. Scassellat, and M. Williamson.

9. "The Cyc Knowledge Server" by CYCORP, available at http://www.cyc.com/products2.html.

9. CYCORP, "CycL Language Features," http://www.cyc.com/cycl.html. 2010.

11.5 "Query by Image and Video Content: The QBIC System" by M. Flickner et al., published in Computer, volume 28, issue 9, pages 23–32, September 1999, and also accessible online at http://wwwqbic.almaden.ibm.com/.

11. "Mental Models" by P. N. Johnson-Laird, published in 1983 by Cambridge University Press in Cambridge, MA.

13, "Semiotics and Agents for Integrating and Navigating through Multimedia Representations of Concepts" (IS&T/SPIE-2000), pp. 120-131, San Jose, California, Jan. 2000, by D. W. Joyce, P. H. Lewis, R. H. Tansley, M. R. Dobie, and W. Hall.

S. M. Kosslyn's 1980 book "Image and Mind" was published by Harvard University Press in Cambridge, Massachusetts.

"The Dimensions of Context Space" by D. Lenat, October 1998, http://www.cyc.com/context-space.doc.

16. In "Towards multimedia thesaurus support for media-based navigation," P. Lewis, H. Davis, M. Dobie, and W. Hall (1996) presented their work at the Image Databases and Multimedia Search (IDB-MMS-1996) conference in Amsterdam. The paper spans pages 83 to 90.

17. "A Texture Thesaurus for Browsing Large Aerial Photographs" by W. Y. Ma and B. S. Manjunath, published in the May 1998 issue of the Journal of the American Society for Information Science (JASIS), is available online and spans pages 633-648.

19. "Semiotic Modeling and Situation Analysis: An Introduction" by A. Meystel, published by AdRem in 1995, Bala Cynwyd, PA.

19. "WordNet: A Lexical Database for English" by G. A. Miller, published in November 1995 in Communication of the ACM, Vol. 38, No. 11, pp. 39-41.

20. In P. Winston's The Psychology of Computer Vision, M. Minsky wrote "A Framework from Representing Knowledge" on pages 211-277. The book was published by McGraw-Hill in 1975 in New York.

"MPEG-7 Multimedia Description Schemes XM (v4.0)" was published in July 2000 in Beijing, China by the MPEG Multimedia Description Schemes Group and was approved by the ISO/

IEC JTC1/SC29/WG11 MPEG00/N3465.

"MPEG-7 Multimedia Description Schemes WD (v4.0)" was published in July 2000 in Beijing, China by the MPEG Multimedia Description Schemes Group and was approved by the ISO/IEC JTC1/SC29/WG11 MPEG00/N3466.

"MPEG-7: Context, Objectives and Technical Roadmap, V.12" was published in July 1999 in Vancouver by the ISO/IEC JTC1/SC29/WG11 MPEG99/N2861.

S. Paek, V. Hatzivassiloglou, A. Jaimes, B. H. Schiffman, S.-F. Chang, and K. R. McKeown presented their work at the ACM SIGIR Workshop on Multimedia Indexing and Retrieval (ACM SIGIR-1999) in August 1999 in Berkeley, California.

25. "Semantic Memory" by M. R. Quillian, in Semantic Information Processing (ed. by M. Minsky), New York: MIT Press, 1968.

26. "Relevance Feedback Techniques in Interactive Content-Based Image Retrieval" by Y. Rui, T. S. Huang, and S. Mehrotra was published in January 1998 in San Jose, California, USA, at the Proceedings of the Conference on Storage and Retrieval of Image and Video Databases VI (IS&T/SPIE-1998).

27. "Representation of Knowledge" by D. E. Rumelhart and D. A. Norman, in Issues in Cognitive Modeling (eds. A. M. Aitkenhead and J. M. Slack), Lawrence Erlbaum Associates, London, 1985.

Prentice Hall, Englewood Cliffs, NJ, 1995. 28. "Artificial Intelligence: A Modern Approach" by S. J. Russell and P. Norvig.

"Quantitative Assessment of Image Retrieval Effectiveness," by J. R. Smith, forthcoming in the Journal of Information Access, is item 29.

30. "Conceptual Modeling of Audio-Visual Content" by J. R. Smith and A. B. Benitez was published in July 2000 in New York, NY, USA, at the International Conference On Multimedia and Expo (ICME-2000).

This information may be found at http://www.ctr.columbia.edu/VisualSEEk/, as well as in the paper "VisualSEEk: A Fully Automated Content-Based Image Query System" (J. R. Smith and S.-F. Chang, 1996).

32. "SaFe: A General Framework for Integrated Spatial and Feature Image Search" (IEEE 1997 Workshop on Multimedia Signal Processing, authored by J. R. Smith and S.-F. Chang), 1997.

33. "Multimedia Search: An Authoring Perspective" by S. W. Smoliar, J. D. Baker, T. Nakayama, and L. Wilcox was published in the Proceedings of the First International Workshop on Image Databases and Multimedia Search (IAPR-1996), which took place in August 1996 in Amsterdam, The Netherlands, and was printed on pages 1-8.

34. "Indoor-Outdoor Image Classification" by M. Szummer and R. Picard, presented at the 1998 IEEE International Workshop on Content-Based Access to Image and Video Databases, that was held in Bombay, India, in connection with ICCV'98.

Rhys Tansley's "The Multimedia Thesaurus: Adding A Semantic Layer to Multimedia Information" was published in August 2000 as a doctoral thesis in computer science from the University of Southampton in the United Kingdom.

36. In their paper "On Image Classification: City vs. Landscape," Vailaya, Jain, and Zhang (in press) discussed this topic at the 1998 IEEE Workshop on Content-Based Access of Image and Video Libraries in Santa Barbara, California.