New Frontiers for Intelligent Content-Based Retrieval

Ana B. Benitez [*] [a], John R. Smith [b]

[a] Dept. of Electrical Engineering, Columbia University, New York, NY 10027

[b] IBM T. J. Watson Research Center, New York, NY 10532

ABSTRACT

In this paper, we examine emerging frontiers in the evolution of content-based retrieval systems that rely on an intelligent infrastructure. Here, we refer to intelligence as the capabilities of the systems to build and maintain situational or world models, utilize dynamic knowledge representations, exploit context, and leverage advanced reasoning and learning capabilities. We argue that these elements are essential to producing effective systems for retrieving audio-visual content at semantic levels matching those of human perception and cognition. In this paper, we review relevant research on the understanding of human intelligence and construction of intelligent systems in the fields of cognitive psychology, artificial intelligence, semiotics, and computer vision. We also discuss how some of the principal ideas from these fields lead to new opportunities and capabilities for content-based retrieval systems. Finally, we describe some of our efforts in these directions. In particular, we present MediaNet, a multimedia knowledge presentation framework, and some MPEG-7 description tools that facilitate and enable intelligent content-based retrieval.

Keywords: Intelligent content-based retrieval, cognitive psychology, artificial intelligence, semiotics, dynamic knowledge, context, reasoning, learning, MediaNet, MPEG-7

1. NEED FOR INTELLIGENCE IN CONTENT-BASED RETRIEVAL

Recent research on the analysis of audio-visual content has enabled a first-generation of multimedia information systems that support content-based retrieval, automatic but constrained classification of objects and scenes, and simple mechanisms for learning such as by relevance feedback from users. These developments represent significant advances from a complete reliance on textual keywords for indexing and retrieval; however, we have yet to see these capabilities substantially differentiate operational multimedia information systems. In many ways, the development of these technologies has been a step in the right direction, but we are just not there yet.

We believe the next-generation multimedia information systems will be marked by the full introduction of intelligence into the systems. This development will be driven by work from fields such as cognitive psychology, artificial intelligence, semiotics, and computer vision that will greatly enhance the initial contributions from signal processing and pattern analysis. We believe that future content-based retrieval systems need to be capable of communicating with the user and understanding of audio-visual content at a higher semantic level in order to be truly effective. Of course, this presents us with the many known problems of object and scene recognition, complex scene understanding, and advanced reasoning and learning, among others.

2. RECENT WORK ON CONTENT-BASED RETRIEVAL

Numerous content-based retrieval (CBR) systems have explored the possibilities of indexing image and video content by using low-level visual features (for example, see Virage [2], QBIC [11], and VisualSeek [31]). These systems work by (1) automatically extracting features directly from the visual data; (2) indexing the extracted descriptors for fast access; and (3) querying and matching descriptors for the retrieval of the visual data. Beyond these basic capabilities, there has been an effort to support relevance feedback to refine queries and learn through examples what the user may be looking for [26].

[*] Correspondence: Email: ana@ee.columbia.edu; WWW: http://www.ee.columbia.edu/~ana/; Telephone: +1-212-854-7473; Fax: +1-212-932-9421

More recently, there has been focused effort on automatically producing certain semantic labels that could contribute significantly to retrieving visual data. For example, recent work has focused on portrait vs. landscape detection, indoor vs. outdoor classification, city vs. landscape classification, sunset vs. forest classification [24, 34, 36], and other attempts to answer basic questions of who, what, when, and where about the visual content. Most of the approaches rely on traditional machine learning techniques to produce semantic labels, and some degree of success has been reached for various constrained and sometimes skewed test sets. However, these efforts represent only a small initial step towards achieving real understanding of the content.

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

The primary objectives of the field of artificial intelligence [28], or AI, are to understand intelligent entities and to construct intelligent systems. Approaches on artificial intelligence can be divided along two main dimensions: behavior vs. thought; and human performance vs. rationality - i.e. ideal concept of intelligence. Another important distinction is symbolic vs. non-symbolic or intermediate approaches [6, 7].

Symbolic approaches follow the physical symbol system hypothesis, which maintains that any physical symbol system has the necessary and sufficient means for general intelligent behavior [28]. A physical symbol system involves inter-related physical patterns called symbols and mechanisms for managing these symbols. Symbols are defined as abstractions of perceptions and actions of the real world. For example, the word “car” is a symbol of the notion of a real object car. AI approaches to rational behavior and thinking are mainly based on physical symbol systems. However, symbolic, non-symbolic, and, more recently, intermediate approaches have been investigated for modeling human behavior and thinking.

The important contributions of the symbolic approaches in AI have been the computerization of logics and reasoning systems, and the development of computer-based knowledge representation models. Non-symbolic and intermediate approaches have resulted in more complex reactive systems based on direct or processed percepts. The objective has often been to enable robots and other intelligent systems to understand the real world well enough to navigate and reason about their environments and automatically achieve certain goals, in-line with previously mentioned human cognitive processes. Logics, knowledge representation models, and reactive systems shall be discussed in the following subsections.

3.2.1. Logics

Logics aim at emulating the laws of thought by providing a mechanism to represent statements of the world – the representation language - and a set of rules to deduce new statements from previous ones – the proof theory. The representation language is defined by its syntax and semantics, which specify the structure and the meaning of the statements, respectively. Different logics make different commitments on what exits in the world (e.g. facts) and on the beliefs about the statements (e.g. true/false/unknown). The most widely used and understood logic is First-Order Logic (FOL), also known as First-Order Predicate Logic (FOPL), which assumes (1) the existence of facts, objects (individual entities), and relations among objects; and (2) the beliefs of true, false, and unknown. Logics represent propositional statements of the world, in other words, statements of verbal or textual nature.

3.2.2. Knowledge Representation Models

Semantic networks, frames, and scripts are schemes that have been proven to be effective in knowledge representation. It is accepted that knowledge bases using these schemes could be expressed using any logic such as FOL. Semantic networks [25] use nodes to represent objects, concepts, or situations; and arcs to represent relationships between nodes (e.g. the state "Bill is a person" could be represented by the chain: Bill Clinton Node - Is-A Arc – Person Node). In spite of their simplicity and support for modular inheritance, semantic networks suffer from limited expressiveness as they can not represent negation or disjunction, among others.

Frames and scripts distinguish between stereotypical and instance situations. A frame [20] is a network of nodes and relations whose higher levels represent attributes that are always true about the situation and whose lower levels contain information about specific instances of the situation. Default values can be specified and cancelled for frame attributes. A script is similar to a frame with additional information about the expected sequence of events; and the goals and the plans of the actors involved. Criticisms of frames and scripts are that definitions are indeed quite important; that cancellation and default values cause problems; and that it is impossible to consider all the relevant aspects of common situations.

The Cyc knowledge server [9] aims at representing and reasoning on everyday life knowledge, i.e. an encyclopedia. The Cyc representation language [10], CycL, is an extension of FOL to handle equality and default reasoning, among others. Facts about the world are grouped and positioned in independent contexts to enable efficient entry and access of the knowledge. Contexts are defined along 12 almost independent dimensions [15] such as absolute time (e.g. on January 1, 2000), type of time (e.g., at night), absolute place (e.g, in New York City), type of place (e.g., outdoors), culture (e.g., Catholic), and topic (e.g., about space). Cyc has been used to enhance the retrieval of photos by processing the textual captions associated with the photos with Cyc natural-language processor.

Other examples of knowledge representation frameworks are text-based frameworks such as WordNet and multimedia-based frameworks such as the Multimedia Thesaurus and visual pattern libraries. WordNet [19] is an electronic lexical system that organizes English words into sets of synonyms, each representing a lexicalized concept, and links them with semantic relationships. The semantic relationships incorporated by WordNet are synonymy, antonymy, hypernymy/hyponymy, meronymy/holonymy, entailment, and troponymy whose definitions are listed in Table 1 with examples. WordNet has been used in the retrieval of text documents and images with associated text annotations to extent keyword queries and the database content [1].

The Multimedia Thesaurus [16, 35] is a network of concepts, relationships between the concepts, and media representations of the concepts. Concepts are abstractions of semantically meaningful objects in the real world and are represented by portions of multimedia materials and associated feature vectors. The relationships between concepts are the typical thesaurus relationships - specialization/specialization, related, and equivalent. In visual pattern libraries such as the texture image thesaurus [17] and the SaFe system [32], concepts represent visual patterns and are defined based on perceptual information such as color and texture features. The Multimedia Thesaurus and the visual pattern libraries have been used to enhance content-based retrieval, browsing, and navigation [16, 17, 32].

3.2.3. Reactive Systems

Non-symbolic approaches to artificial intelligence [6] are purely sensory stimulus-based. These systems act on and response to sensory information often reaching decisions through very complex processes that are often difficult to understand and

justify. Although successful in certain tasks such as navigation, an important difficulty that these systems face is how to transfer their knowledge to other systems because they lack symbols or comparable abstractions to represent knowledge.

The intermediate approach described in [7] aims at generating symbolic-like representations based on sensory stimulus through recursive processes of generalization (i.e. grouping based on similarity) without instantiating actual symbols. This work also proposes four essences of intelligence: the progressive development of the system skills; the social interaction with other systems and humans for learning and perfecting skills; the integration of complementary sensory and motor skill; and the experience of a body and physical coupling [8].

Table 1: Definitions and examples of the semantic relationships in WordNet.

Relationship	Definition	Example
Synonymy	To be similar to	Human *is similar to* homo
Antonymy	To be opposite to	White is *opposite to* Black
Hypernymy / Hyponymy	To be an specialization To be a generalization	Hominid *is an specialization of* Human Human *is a generalization of* Hominid
Meronymy / Holonymy	To be a part, member, or substance of To have a part, member, or substance of	Ship *is a member of* Fleet Martini *has substance* Gin
Entailment	To cause or involve by necessity or as a consequence	Divorce *entails* Marry
Troponymy	To be a manner of	Whisper *is a manner of* Speak

5.3. Semiotics

Semiotics is the science that analyzes signs (and sign systems) and puts them in correspondence with particular meanings (e.g. word "car" and notion of real object car). Examples of sign systems are conversational and musical languages, and artificial logics. For our interests, a more adequate definition of semiotics is provided by situational analysis: "Semiotics is a theoretical field which analyzes and develops formal tools of knowledge acquisition, representation, organization, generation and enhancement, communication, and utilization" [18]. The latter definition is in strong correlation with the well-known decomposition of semiotics into three domains: syntax (sign; "car"), semantic (interpretant; notion of car), and pragmatics (object; real object car); and the Six-Box Diagram for modeling intelligent behavior and thinking: world, sensors, perception, knowledge, decision making (planning and control), and actuators (see Figure 1).

Figure 1 shows the knowledge cycle and the components of semiotics. The world and what happens in the world is encoded by sensors in a symbolic form. The role of perception is to represent the results of sensing in some organized manner using signs (syntax). Through further organization and generalization, this information becomes knowledge. Interpretation of the the knowledge is necessary to enable the process of decision making where the interpretant is created by adding semantics to syntax. Actuation is analogous to the process of generating new knowledge and it is based on the interpretant. The new knowledge arrives in the world and changes in the world, physically and/or conceptually. New objects emerge that can be encoded by sensors closing the cycle.

The principle of multi-resolution is semiotics arises from modeling a unit of intelligence as three cognitive processes applied repeatedly: focusing attention, combinatorial search, and grouping (or generalization). First, attention is focused on a subset of the available data. Then, different combinations of this data are generated based on some similarity criteria. The combination or grouping providing the best results generates data at the next level. The multi-resolution framework is in accordance with Gödel's theorem of incompleteness, which evokes the need for an external body of knowledge, for example, a meta-language, to interpret some of the statements that cannot be proven within a particular language. The result of this theorem is a multi-resolution hierarchy of languages.

There has been some recent work connecting the fields of semiotics and multimedia information systems. Smoliar et al. [33] describes some of the implications to multimedia search from the point of view of writing and reading multimedia signs. Multimedia material such as images and words are considered to signify notions of objects in the world (e.g., an image of a carrot and the word "carrot" signify the notion of carrot); and search, fundamental for the processes of reading and writing. Joyce et al. [13] proposes a semiotics framework to integrate high-level metadata (e.g. "carrot") and low-level metadata (e.g. color histogram extracted from the image of a carrot) by formally adding a second representation level to [33]. This level

consists of features extracted from the multimedia material acting as signs of the multimedia material itself (see Figure 2). Textual and non-textual features signs are identified as high-level and low-level metadata, respectively. The link between the two is established with the Multimedia Thesaurus [16, 35] and neural-network classification agents [13]. Del Bimbo [5] applies the semiotics idea of producing meaning at two levels, the narrative level and discourse level, to automatically annotate and retrieve videos of commercials. The narrative level includes basic signs and the results of sign combinations; the discourse level describes how to use narrative elements to create a story.

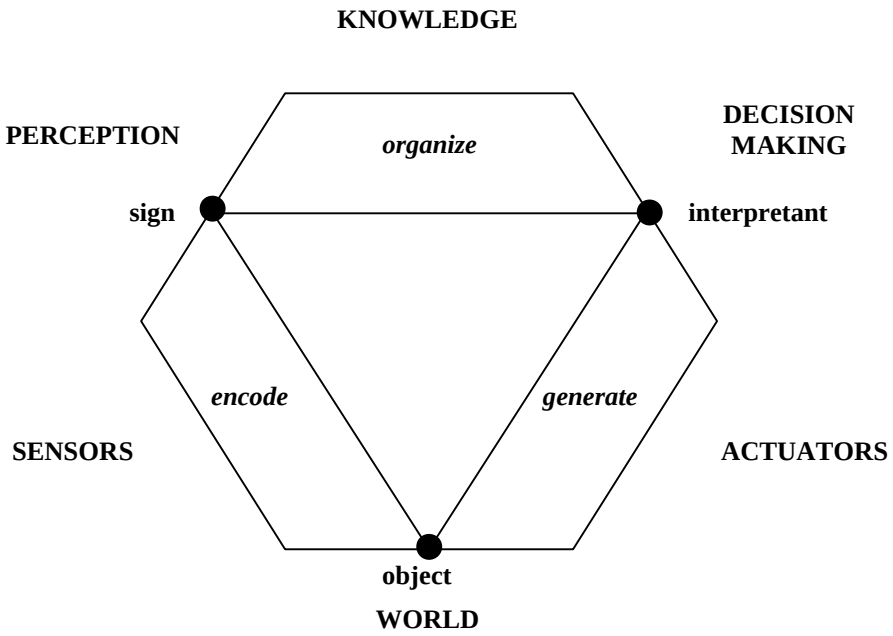

Figure 1: Relationship between the three domains of semiotics and the six-box diagram for modeling intelligent behavior and thinking.

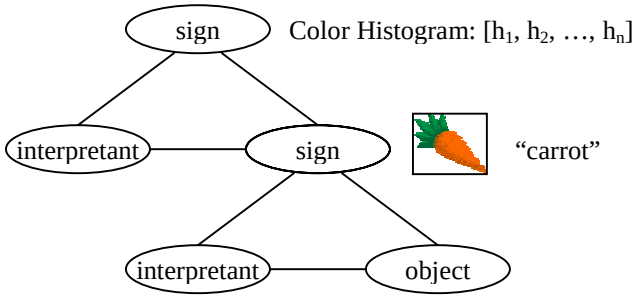

Figure 2: Semiotic framework for multimedia and features extracted from multimedia [13].

5.4. Computer Vision

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.

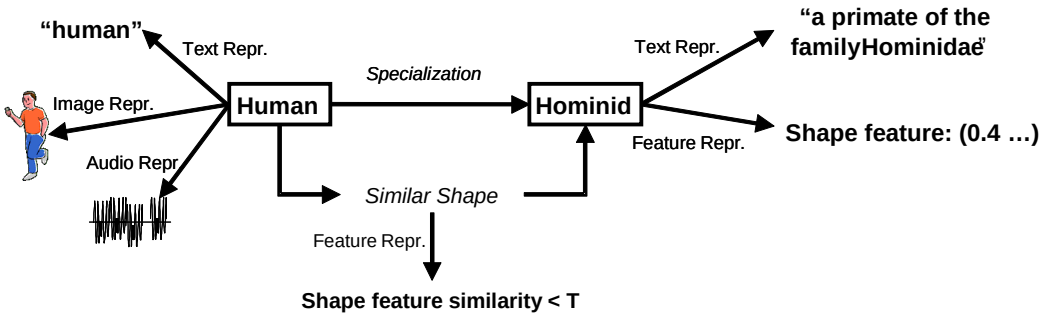

Figure 3: Example of a MediaNet that illustrates the concept Human and Hominid (boxes) with multiple text, image, audio, and shape feature representations. The two concepts are related by a semantic relationship, *Specialization*, and a perceptual relationship, *Similar Shape*, with an associated shape 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.

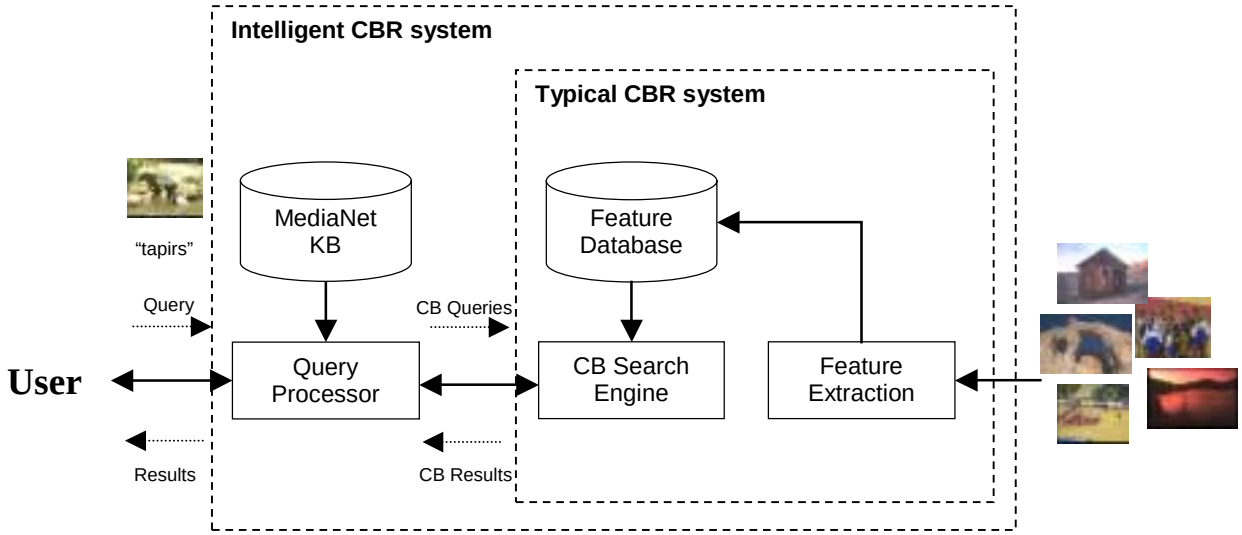

Figure 4: Components of the implemented intelligent content-based retrieval system that extends a typical content-based retrieval system with a MediaNet knowledge base and a query processor.

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.

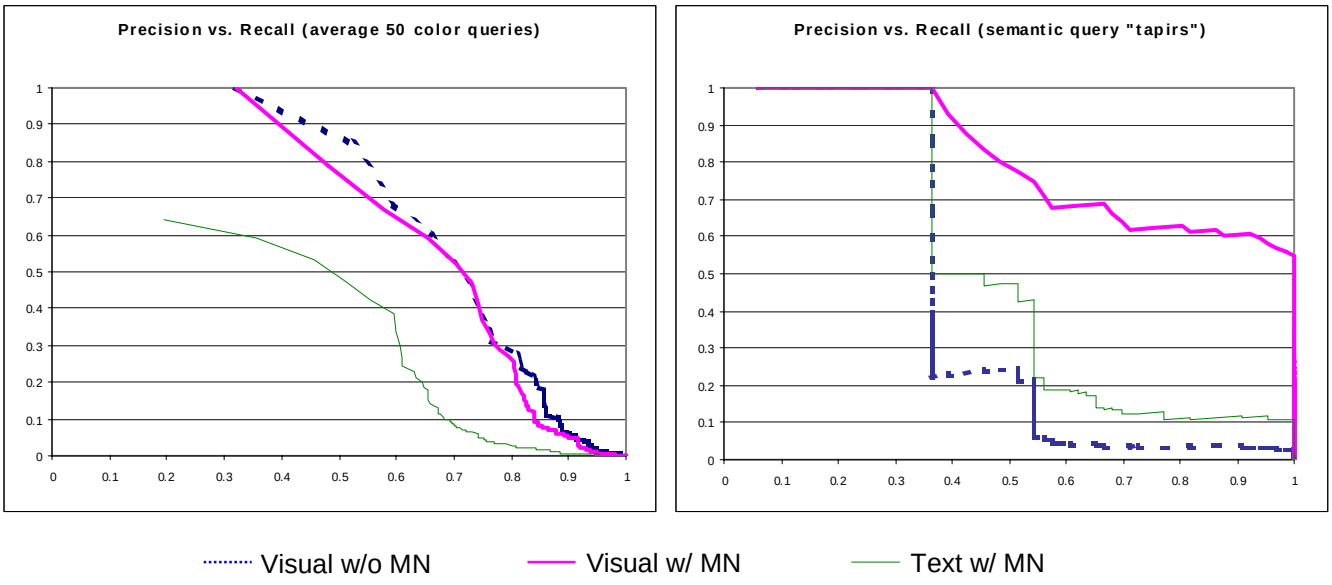

Figure 5: Average precision and recall for the 50 MPEG-7 color queries and the semantic query "tapirs" using color histogram. "Visual w/o MN" corresponds to the typical content-based retrieval system that does not use MediaNet; "Visual w/ MN" to the intelligent content-based retrieval system with image queries; "Text w/ MN" to the intelligent content-based retrieval system using the keywords as queries.

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.

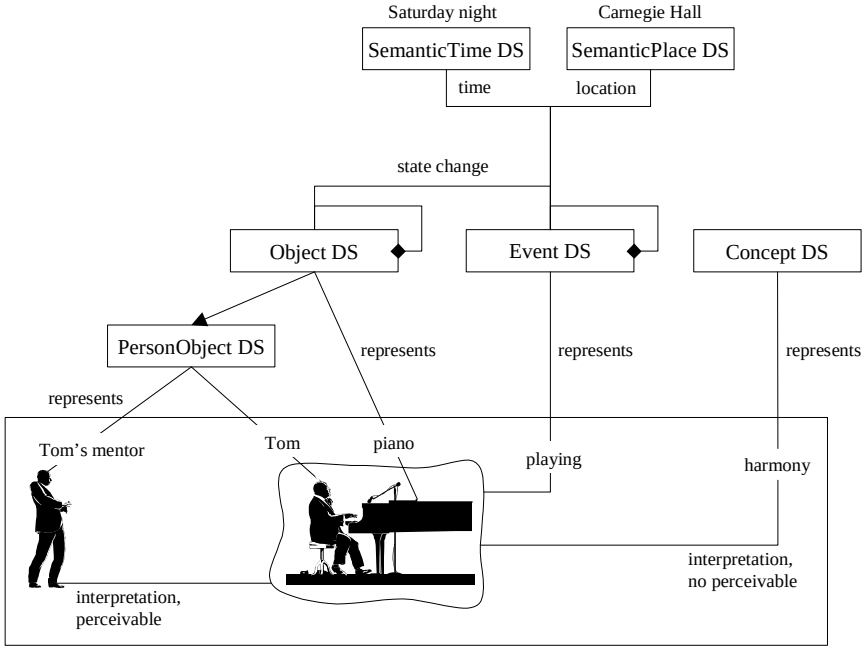

Figure 6: Semantic description of the narrative world depicted in piece of audio-visual content. The description states that Tom is playing the piano on Saturday night at Carnegie Hall. This event is interpreted as being harmonic and a tribute to Tom’s mentor.

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

This paper has examined research on the understanding of intelligent entities and the construction of intelligent systems in the fields of psychology, artificial intelligence, semiotics, and computer vision to enable intelligent content-based retrieval systems. We refer to intelligence as the capabilities of the systems to build and maintain situational or world models, utilize dynamic knowledge representations, exploit context, and leverage advanced reasoning and learning capabilities. We have discussed the implications and opportunities for intelligent content-based retrieval systems. Some of our efforts in these directions have focused on the development of MediaNet, a multimedia knowledge presentation framework, and MPEG-7 description tools that facilitate and enable intelligent content-based retrieval, that we have also presented in this paper.

REFERENCES

1. Y. A. Aslandogan, C. Their, C. T. Yu, and N. Rishe, "Using Semantic Contents and WordNet in Image Retrieval", *Proc. of the 20th It.l ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 286-295, 1997.
2. J. R. Bach et al., "Virage Image Search Engine: An Open Framework for Image Management", *Proceeding of Conference on Storage and Retrieval for Image and Video Databases IV (IS&T/SPIE-1996)*, San Jose, California, 1996.
3. D. H. Ballard and C. M. Brown, *"Computer Vision"*, PRENTICE-HALL INC., Englewood Cliffs, NJ, 1982.
4. A. B. Benitez, J. R. Smith, and S.-F. Chang, "MediaNet: A Multimedia Information Network for Knowledge Representation", *Proceedings of the Conference on Internet Multimedia Management Systems (IS&T/SPIE-2000)*, Boston, MA, November 2000.
5. A. Del Bimbo, "Expressive Semantics for Automatic Annotation and Retrieval of Video Streams", *Proceedings of International Conference On Multimedia and Expo (ICME-2000)*, New York, NY, July 2000.
6. R. A. Brooks, "Intelligence Without Reason", MIT AI Lab Memo 1293, April 1991.
7. R. A. Brooks and L. A. Stein, "Building Brains for Bodies", *Autonomous Robots*, Vol. 1, No. 1, pp. 7-25, Nov. 1994.
8. R. A. Brooks, C. Breazeal, R. Irie, C. Kemp, M. Marjanovic, B. Scassellat, and M. Williamson, "Alternate Essences of Intelligence", appeared in AAAI-98.
9. CYCORP, "The Cyc Knowledge Server", http://www.cyc.com/products2.html.
10. CYCORP, "Features of the CycL Language", http://www.cyc.com/cycl.html.
11. M. Flickner et al., "Query by Image and Video Content: The QBIC System", *Computer*, Vol. 28, No. 9, pp. 23-32, Sep. 19995; also available at http://wwwqbic.almaden.ibm.com/.
12. P. N. Johnson-Laird, *"Mental Models"*, Cambridge University Press, Cambridge, MA, 1983.
13. D. W. Joyce, P. H. Lewis, R. H. Tansley, M. R. Dobie, and W. Hall, "Semiotics and Agents for Integrating and Navigating through Multimedia Representations of Concepts", *Proceedings of the Conference on Storage and Retrieval for Media Databases (IS&T/SPIE-2000)*, pp. 120-131, San Jose, California, Jan. 2000.
14. S. M. Kosslyn, *"Image and Mind"*, Harvard University Press, Cambridge, MA, 1980.
15. D. Lenat, "The Dimensions of Context Space", http://www.cyc.com/context-space.doc, Oct. 1998.

16. P. Lewis, H. Davis, M. Dobie, and W. Hall, "Towards multimedia thesaurus support for media-based navigation", *Image Databases and Multimedia Search: Proceedings of the First International Workshop, (IDB-MMS-1996)*, pp. 83-90, Amsterdam, 1996.
17. W. Y. Ma and B. S. Manjunath, "A Texture Thesaurus for Browsing Large Aerial Photographs", *Journal of the American Society for Information Science (JASIS)*, pp. 633-648, vol. 49, No. 7, May 1998.
18. A. Meystel, *"Semiotic Modeling and Situation Analysis: An Introduction"*, AdRem, Bala Cynwyd, PA, 1995.
19. G. A. Miller, "WordNet: A Lexical Database for English", *Communication of the ACM*, Vol. 38, No. 11, pp. 39-41, Nov. 1995.
20. M. Minsky, "A Framework from Representing Knowledge", The Psychology of Computer Vision, P. Winston (ed), pp. 211-277, McGraw-Hill, New York, NY, 1975.
21. MPEG Multimedia Description Schemes Group, "MPEG-7 Multimedia Description Schemes XM (v4.0)", ISO/IEC JTC1/SC29/WG11 MPEG00/N3465, Beijing, CN, July 2000.
22. MPEG Multimedia Description Schemes Group, "MPEG-7 Multimedia Description Schemes WD (v4.0)", ISO/IEC JTC1/SC29/WG11 MPEG00/N3466, Beijing, CN, July 2000.
23. MPEG Requirements Group, "MPEG-7: Context, Objectives and Technical Roadmap, V.12", ISO/IEC JTC1/SC29/WG11 MPEG99/N2861, Vancouver, July 1999.
24. S. Paek, C. L. Sable, V. Hatzivassiloglou, A. Jaimes, B. H. Schiffman, S.-F. Chang, K. R. McKeown, "Integration of Visual and Text based Approaches for the Content Labeling and Classification of Photographs", *ACM SIGIR Workshop on Multimedia Indexing and Retrieval (ACM SIGIR-1999)*, Berkeley, CA, Aug. 1999.
25. M. R. Quillian, *"Semantic Memory"*, Semantic Information Processing, M. Minsky (ed), MIT Press, Cambridge, MA, 1968.
26. Y. Rui, T. S. Huang, and S. Mehrotra, "Relevance Feedback Techniques in Interactive Content-Based Image Retrieval", *Proceedings of the Conference on Storage and Retrieval of Image and Video Databases VI, (IS&T/SPIE-1998)*, San Jose, California, Jan. 1998.
27. D. E. Rumelhart and D. A. Norman, "Representation of Knowledge", in A. M. Aitkenhead & J. M. Slack (Eds.), *Issues in Cognitive Modeling*, Lawrence Erlbaum Associates, London, 1985.
28. S. J. Russell and P. Norvig, *"Artificial Intelligence: A Modern Approach"*, Prentice Hall, Englewood Cliffs, NJ, 1995.
29. J. R. Smith, "Quantitative Assessment of Image Retrieval Effectiveness", To appear in *Journal of Information Access*.
30. J. R. Smith and A. B. Benitez, "Conceptual Modeling of Audio-Visual Content", *Proceedings of the International Conference On Multimedia and Expo (ICME-2000)*, New York, NY, July 2000.
31. J. R. Smith and S.-F. Chang, "VisualSEEk: A Fully Automated Content-Based Image Query System", *Proceedings of the ACM Conference Multimedia*, New York, 1996; also available at http://www.ctr.columbia.edu/VisualSEEk/.
32. J. R. Smith and S.-F. Chang, "SaFe: A General Framework for Integrated Spatial and Feature Image Search", *IEEE 1997 Workshop on Multimedia Signal Processing*, 1997.
33. S. W. Smoliar, J. D. Baker, T. Nakayama, and L. Wilcox, "Multimedia Search: An Authoring Perspective", *Proceedings of the First International Workshop on Image Databases and Multimedia Search (IAPR-1996)*, pp. 1-8, Amsterdam, The Netherlands, Aug. 1996.
34. M. Szummer and R. Picard, "Indoor-Outdoor Image Classification", *IEEE International Workshop in Content-Based Access to Image and Video Databases*, in conjunction with ICCV'98, Bombay, India, Jan. 1998.
35. R. Tansley, "The Multimedia Thesaurus: Adding A Semantic Layer to Multimedia Information", Ph. D. Thesis, Computer Science, University of Southampton, UK, Aug. 2000.
36. A. Vailaya, A. Jain, and H.J. Zhang, "On Image Classification: City vs. Landscape", *IEEE Workshop on Content-Based Access of Image and Video Libraries*, Santa Barbara, CA, June 1998.