MediaNet: A Multimedia Information Network for Knowledge Representation

Ana B. Benitez [*] [ab], John R. Smith [a], Shih-Fu Chang [b]

[a] IBM T. J. Watson Research Center, New York, NY 10532
[b] Dept. of Electrical Engineering, Columbia University, New York, NY 10027

ABSTRACT

In this paper, we present MediaNet, which is a knowledge representation framework that uses multimedia content for representing semantic and perceptual information. The main components of MediaNet include conceptual entities, which correspond to real world objects, and relationships among concepts. MediaNet allows the concepts and relationships to be defined or exemplified by multimedia content such as images, video, audio, graphics, and text. MediaNet models the traditional relationship types such as generalization and aggregation but adds additional functionality by modeling perceptual relationships based on feature similarity. For example, MediaNet allows a concept such as "car" to be defined as a type of a "transportation vehicle", but which is further defined and illustrated through example images, videos and sounds of cars. In constructing the MediaNet framework, we have built on the basic principles of semiotics and semantic networks in addition to utilizing the audio-visual content description framework being developed as part of the MPEG-7 multimedia content description standard.

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. In particular, we have found that MediaNet can improve the performance of multimedia retrieval applications by using query expansion, refinement and translation across multiple content modalities. In this paper, we report on experiments that use MediaNet in searching for images. We construct the MediaNet knowledge base using both WordNet and an image network built from multiple example images and extracted color and texture descriptors. Initial experimental results demonstrate improved retrieval effectiveness using MediaNet in a content-based retrieval system.

Keywords: MediaNet, concept, multimedia, content-based retrieval, MPEG-7, WordNet, semiotics, semantic networks, intelligence

1. INTRODUCTION

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 [16]. On the other hand, text or words in a language can be thought of as symbols for the real world entities. The mapping from the content level to the symbolic level by computer is quite limited and far from reaching human performance. As a result, in order to deal effectively with audio-visual material, it is necessary to model real world concepts and their relationships at both the symbolic and perceptual levels.

In order to address the problem of representing real world concepts using semantics and perceptual features, we propose the MediaNet multimedia knowledge representation framework. MediaNet represents the real world using concepts and relationships that are defined and exemplified using multiple media. MediaNet can be used to facilitate the extraction of knowledge from multimedia material and improve the performance of multimedia searching and filtering applications. In MediaNet, concepts represent real world entities. Furthermore, relationships can be conceptual (e.g., Is-A-Subtype-Of) and perceptual (e.g., Is-Similar-To). The framework offers functionality similar to that of a dictionary or encyclopedia and a

[*] Correspondence: Email: ana@ee.columbia.edu; WWW: http://www.ee.columbia.edu/~ana/; Telephone: +1-212-316-9136; Fax: +1-212-932-9421

thesaurus by defining, describing and illustrating concepts, but also by denoting the similarity of concepts at the semantic and perceptual levels.

1.1. Related Work

Previous work has focused on the development of multimedia knowledge bases for information retrieval such as the multimedia thesaurus (MMT) [19], a central component of the MAVIS 2 system [3], and the texture image thesaurus [7]. The multimedia thesaurus provides concepts, which are abstract entities of the real world objects, semantic relationships such a generalization and specialization, and media representations of the concepts, which are portions of multimedia materials and associated features vectors. MediaNet extends this notion of relationships to include perceptual relationships that can also be exemplified and defined using audio-visual content. Furthermore, MMT treats semantic objects and perceptual information quite differently. MMT defines concepts that correspond to high-level, semantically meaningful objects in the real world with names in a language ("car" and "man") [5, 20]. However, this exclude concepts that represent patterns based on perceptual information that are not named, such as the texture patterns in the texture image thesaurus [7].

MediaNet extends the current knowledge representation frameworks by including multimedia information. 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 feature levels. In particular, we have found that MediaNet can improve the performance of multimedia retrieval applications by using query expansion, refinement and translation across multiple content modalities. In this paper, we report on experiments that use MediaNet in searching for images. We construct the MediaNet knowledge base using both WordNet and an image network built from multiple example images and extracted color and texture descriptors. Initial experimental results demonstrate improved retrieval effectiveness using MediaNet in a content-based retrieval system; however, more extensive experiments are needed.

1.2. Outline

This paper is organized as follows. In section 2, we describe the main components of MediaNet and their correspondence to principles in semiotics, semantic networks in AI, and MPEG-7 description tools. Section 3 describes the implementation of an extended content-based retrieval system, which uses the MediaNet framework. In particular, it focuses on the construction and the use of the MediaNet knowledge base. Section 4 reports on the experiments that compare the performance of the extended content-based retrieval system to a typical content-based retrieval system. Finally, section 5 concludes with a summary and open issues.

2. MEDIANET

MediaNet is a semantic network for representing real world knowledge through both symbolic and audio-visual information such as images, video, audio, graphics and text, including extracted feature descriptors. In MediaNet, real world entities are represented using concepts and relationships among concepts, which are defined and exemplified by multimedia material. For example, the concept Human can be represented by the words "human", "man", and "homo", the image of a human, and the sound recording of a human talking, and can be related to the concept Hominid by a Is-A-Subtype-Of relationship; the concept Hominid can be represented by the word "hominid" and the text definition "a primate of the family Hominidae" [+]. A graphical representation of previous example is shown in Figure 1. In the following sections, we describe the main components of MediaNet and how they map to work on multiple disciplines such as semiotics, semantic networks in AI, and MPEG-7 description tools.

1.3. Concepts

Concepts are the basic units for knowledge representation in MediaNet. Concepts refer to semantic notions of objects in the world. Objects are any elements that exist in the world such as inanimate objects, living entities, events, and properties. Examples of concepts are Human (see Figure 1) and Car that refer to any living entity human and any inanimate object car, respectively; Wedding is the concept of an event and Blue, the concept of a property. Concepts can refer to classes of objects in the world such as Car; unique and identified objects such as Ronald Reagan; and abstract objects with no physical presence in the world such as Freedom.

[+] The text definitions of words in this paper were taken from the electronic lexical system WordNet [9].

It is important to point out the differences between words in a language and concepts. In the previous examples, concepts were usually named with one word; however, we humans may have no words to designate some concepts, more than one word to designate the same concept, or no words to uniquely designate a concept. An example of the first case is the unknown texture of a specific piece of fabric. The second case corresponds to synonyms, i.e., words having the same or nearly the same meaning or sense in a language, as "human" and "man" (see Human concept in Figure 1). Finally, the third case corresponds to polysemy, i.e., a word having multiple meanings in a language, as "studio" which can be a "the workroom or atelier of an artist", "a room or set of rooms specially equipped for broadcasting radio or television programs, etc.", and "an apartment consisting of one main room, a kitchen, and a bathroom".

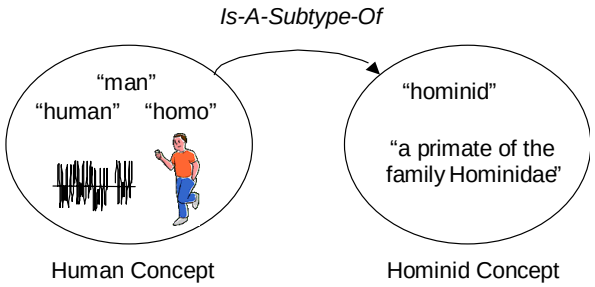

Figure 1: Concepts Human and Hominid with multiple media representations and related by an Is-A-Subtype-Of relationship.

1.4. Media Representations of Concepts

Concepts, which refer to objects in the world, can be illustrated and defined by multimedia content such as text, image, and video, among others. As an example, the concept Human is exemplified by the words "human", "man", and "homo", the image of a human, and the sound recording of a human in Figure 1.

Media representations are not necessarily whole multimedia materials such as an entire image; they can be sections of multimedia material and have associated features extracted from the media. A media representation of the concept Human can be a region of an image that contains an object human and the value of the contour shape feature for that region. Concept could also be illustrated by feature values with no multimedia material. As an example, the concept Human can be presented by the value of a contour shape feature. The value of the contour shape feature may be the average of the contour shape feature values for a set of image regions depicting a human; however, the image data is not included in the media representation only the value of the contour shape feature. Although, the image media was used to exemplify the fact that a media representation can be any section of multimedia material and/or features extracted from the multimedia material; the same applies to any other media such as text, audio, and video. Some media representations may be more or less relevant or applicable to concepts. Some example follow. Cars can be of many colors; therefore, a color feature will not be very relevant for the concept Car. Audio-based representations do not apply to the concept Blue. A relevance factor could be assigned to each representation of a concept.

We shall provide some more examples of media representations of concepts now. The concept Car can have the following media representations: the word "car", the text definition "an automobile", an image depicting a car together with shape features extracted from it, and the sound recording of a running car. The concept Blue can have the following representations: the English word "blue", the Spanish word "azul", the text definition "the pure color of a clear sky; the primary color between green and violet in the visible spectrum", and the value of the color histogram corresponding to blue. Text representations may be in different languages.

1.5. Relationships Among Concepts

Concepts can be related to other concepts by semantic and perceptual relationships. Semantic relationships relate concepts based on their meaning. All the semantic relationships in the lexical system WordNet [9] except for synonymy apply to

concepts; these are listed in Table 1 together with definitions and examples. Antonymy is a binary, symmetric relationship among concepts; the rest of the relationships in Table 1 are binary and transitive. There is usually one hypernym per concept so this relationship organizes the concepts into a hierarchical structure.

Table 1: Examples of semantic relationships with definitions and examples.

Relationship	Definition	Example
Antonymy	To be opposite to	White Is-Opposite-To Black
Hypernymy/ Hyponymy	To be a super type of To be a sub type of	Hominid Is-A-Supertype-Of Human Human Is-A-Subtype-Of Hominid
Meronymy/ Holonymy	To be a part, member, or substance of To have a part, member, or substance of	Ship Is-Member-Of Fleet Martini Has-Substance Gin
Entailment	To entail or To cause or involve by necessity or as a consequence	Divorce Entails Marry
Troponymy	To be a manner of	Whisper Is-Manner-of Speak

Concepts refer to objects that are perceived by senses. Therefore, concepts can also be related by perceptual relationships, which are based on impressions obtained by the use of senses. Examples of perceptual relationships are visual relationships (e.g., Sky Has-Similar-Color-To Sea) and audio relationships (e.g., Stork Sounds-Similar-To Pigeon). Audio-visual relationships are special because they are recorded in the audio-visual media and, therefore, can have media representations as concepts. These relationships can usually be generated and inferred by automatic tools and expressed as constraints and similarity on audio-visual features. They can also be exemplified by two media representations related with that relationship. Figure 2 exemplifies the relationship "Sounds-Similar-To" using two images.

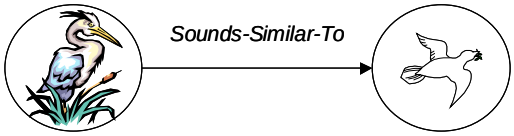

Figure 2: Example of audio relationship Sounds-Similar-To.

1.6. Semiotics View of MediaNet

Semiotics [8] is the science that analyzes signs 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. We will focus on two of the basic principles of semiotics that are most useful for multimedia applications. The first principle is the decomposition of semiotics into three domains: syntax (sign; "car"), semantic (interpretant; notion of car), and pragmatics (object; real object car). The second principle, the multi-resolution, 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. Higher-level thinking skills such as rational reasoning and intuitive thinking are a composition of these basic skills.

In MediaNet, multimedia materials are considered signs of objects in the real world and features extracted from the multimedia material are considered signs of the multimedia material (as proposed in [5]). We identify concepts in our framework with the semantics (interpretant) conferred upon these signs when users interpret them. Although the interpretation of a sign is relative to users and tasks at hand [5], MediaNet aims at recording some interpretations of multimedia and feature signs to enable multimedia applications that intelligently satisfy users' needs to browse and search multimedia content, among others. The information about how users would interpret a sign in one way or another for this or that task could be also included in MediaNet in the form of context information similar to the modality/disposition/epistemology context dimension proposed in [6], which specifies if a fact represents a belief or a desire, and the people who believe in it. The generalization/specialization (or hypernymy/hyponymy) relationship enables the creation of hierarchies of concepts at multiple levels implementing the semiotic principle of multi-resolution.

1.7. MediaNet as a Semantic Network

The primary objectives of the field of artificial intelligence, or AI, are to understand intelligent entities and to construct intelligent systems. Important contributions of AI have been the development of computer-based knowledge representation models such as semantic networks. Semantic networks [14] use nodes to represent objects, concepts, or situations; and arcs to represent relationships between nodes (e.g. the state "Human is a subtype of Homonid" could be represented by the chain: Human Node - Is-A-Subtype-Of Arc – Homonid Node).

The mapping of the components of MediaNet to semantic networks is as follows: concepts and media representations are represented by nodes; arcs link concepts and media representations and are labeled based on the type of representation (e.g., an image representation is linked to a concept with an Image-Representation edge); semantic relationships can be represented by arcs among the concept nodes; and, finally, perceptual relationships that have media representations should be represented by arcs instead of nodes. The example shown in Figure 1 is represented as a semantic network in Figure 3.

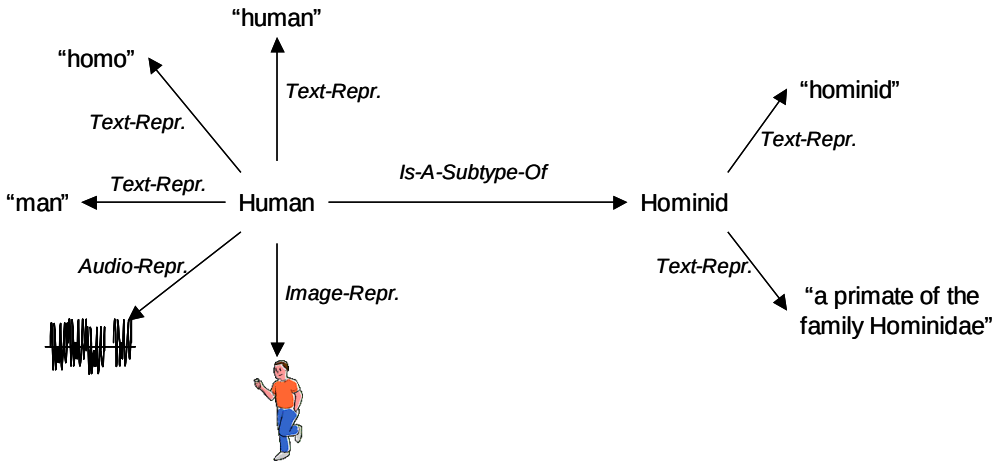

Figure 3: Semantic network corresponding to the MediaNet example shown in Figure 1.

1.8. MPEG-7 Encoding of MediaNet

The MPEG-7 standard [12] aims at standardizing tools for describing the content of multimedia material in order to facilitate a large number of multimedia searching and filtering applications. In this section, we describe and give examples of how MediaNet bases could be encoded using MPEG-7 description tools, which would greatly benefit the exchange and re-use of multimedia knowledge and intelligence among applications.

Relevant MPEG-7 tools for describing MediaNet bases are the semantic and conceptual descriptions schemes, which describe semantic and concepts of the real world as they relate to audio-visual content, and the collection description schemes, which describe collections related to audio-visual content [10,11]. As of the writing of this paper, the collection description schemes are in a more mature state than the semantic and conceptual description schemes so we have picked them for this example. However, it is important to note that these description tools are still under development within the MPEG-7 standard and their specification may change in the near future.

Collection structure description schemes allow describing a multimedia collection as multiple collection clusters (or groupings) of content description elements from the multimedia collection and relationships among these collection clusters. Content description elements can be segments, objects, events, images, and videos, among others. Collection clusters can include text annotations, statistical information (e.g. feature centroids of clusters), and relationships to other collection clusters or content description elements.

The main components of MediaNet could be mapped to the MPEG-7 description tools as follows: a MediaNet knowledge base to a collection structure, each concept to a collection cluster, and each concept relationship to a cluster relationship. The

text representations of a concept could be described as text annotations of collection clusters; while other media representations could be described as cluster elements (e.g., videos and images) and/or cluster statistics (e.g. centroid for concepts). The XML [21] description that encodes the example in Figure 1 is included below. We assume the reader is familiar with the markup language XML.

```
<CollectionStructure id="MediaNet0">
        <CollectionCluster id="ConceptHuman"> <!-- Concept Human -->
                <Text Annotation> <!-- Three textual representations of concept Human -->
                        <FreeTextAnnotation> human </FreeTextAnnotation>
                        <FreeTextAnnotation> homo </FreeTextAnnotation>
                        <FreeTextAnnotation> man </FreeTextAnnotation>
                </Text Annotation>
                <ClusterElements number="2"> <!-- Two media representations, image and audio, of concept Human -->
                        <Segment xsi:type="Image"> <MediaLocator> Human.jpg </MediaLocator> </Segment>
                        <Segment xsi:type="Audio"> <MediaLocator> Human.wav </MediaLocator> </Segment>
                </ClusterElements>
        </CollectionCluster>
        <CollectionCluster id="ConceptHominid"> <!-- Concept Hominid -->
                <Text Annotation> <!-- Two text representations of concept Human -->
                        <FreeTextAnnotation> hominid </FreeTextAnnotation>
                        <FreeTextAnnotation> a primate of the family Hominidae </FreeTextAnnotation>
                </Text Annotation>
        </CollectionCluster>
        <!-- Graph describing Is-A-Subtype-Of relationship from concept Human to concept Hominid -->
        <Graph> <ClusterRelation name="Is-A-Subtype-Of" source="ConceptHuman" target="ConceptHominid"/> </Graph>
</CollectionStructure>
```

3. INTELLIGENT CONTENT-BASED RETRIEVAL SYSTEM

MediaNet extends current knowledge representation frameworks by including multimedia information. 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 feature levels. In this section, we describe the implementation of an extended content-based retrieval system that uses MediaNet to expand, refine, and translate queries across multiple content modalities. In this section, we focus on describing the construction and use of the MediaNet knowledge base in a content-based retrieval system.

Typical content-based retrieval systems index and search image and video content by using low-level visual features (for example, see [1,4,17]). These systems automatically extract low-level features from the visual data, index the extracted descriptors for fast access in a database, and query and match descriptors for the retrieval of the visual data. The extended content-based retrieval system is a typical content-based retrieval system that also includes a MediaNet knowledge base with media representations of concepts and a query processor that uses the MediaNet knowledge base to refine, extend, or translate user queries (see Figure 4).

In the extended content-based retrieval system, we use color and texture features. The color features are color histogram and color coherence; the texture features are wavelet texture and tamura texture. Queries to the CB search engine can only be visual queries as the name of an image in the database or the value for any of the low-level features in the image database. The CB search engine uses weighted Euclidean distance function to obtain distance scores between the query and each image in the database to the query and returns an ordered list of images based on the distance scores. No text annotations are stored in the database or are used in the retrieval.

4.1. Creation of the MediaNet Base

A MediaNet knowledge base could be created manually or automatically using classification, machine learning, and artificial intelligence tools among others. The MediaNet knowledge base for the extended content-based retrieval system was created semi-automatically using existing text annotations for some of the images in the database, the electronic lexical system WordNet, and human input. In this section we describe the procedure followed to construct the MediaNet base.

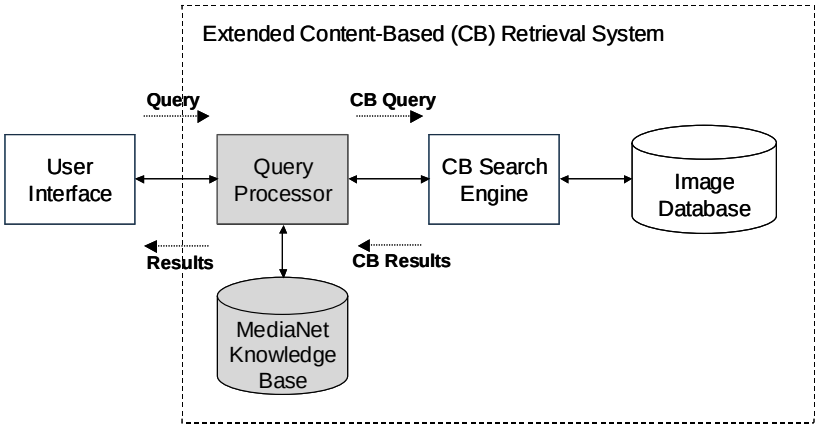

Figure 4: Components of the extended content-based retrieval engine. The new components with respect to typical content-based retrieval engines are shown in gray.

WordNet [9] is an electronic lexical system that organizes English words into set of synonyms, each representing a lexicalized concept. Semantic relationships link the synonym sets. The semantic relationships between words and words senses incorporated by WordNet include synonymy, antonymy, hypernymy/hyponymy, meronymy/holonymy, entailment, and troponymy (these relationships are defined in Table 1 except for synonymy). WordNet contains more than 118,000 different word forms and more than 90,000 word senses. Being available in electronic form, it is one of the most convenient tools for generating concepts, text representations of concepts, and relationships among concepts at the semantic level.

The first step for constructing the MediaNet knowledge was to create concepts and text representations of the concepts using WordNet and human assistance. First, the text annotations were serialized into words that were ordered alphabetically. Dummy words such a prepositions and articles and duplicated words were removed. Then, WordNet was used to generate the senses and the synonyms for each word. A human supervisor selected the correct sense (or senses) of each word in the context of the text annotations and the image data. As an example, for the annotation "singer in studio", the correct sense for the word "studio" was selected as "a room or set of rooms specially equipped for broadcasting radio or television programs, etc.". The supervisor also specified the syntactic category for each word (noun, verb, etc.). A concept was created for each word/sense pair, and was assigned the sense and synonyms provided by WordNet as text representations.

The next step was to generate relationships among concepts. We decided to use the top three relationships listed in Table 1. We used WordNet to automatically generate all of the antonyms, the hypernyms/hyponyms, and the meronyms/holonyms for each concept (i.e., each word-sense pair), automatically parsed the output of WordNet, obtained the relationships among the concepts, and stored them in the MediaNet knowledge base.

Finally, visual representations of concepts were generated automatically using color and texture feature extraction tools. For all the images associated with a concept, we extracted color and texture features and computed the feature centroids (centroid for group of images). The visual representation of each concept was the list of images for the concept with associated feature values, and the feature centroids.

For each application, the list of concepts and relationships in the MediaNet knowledge base should be representative of the content in the database and the goal of the application task. We used the textual annotations already available for the images, which were quite brief and general. More specific and representative text annotations could have been produced and used to construct a more optimized knowledge base. The process of generating the concepts from the words in the textual annotations could be automated by processing the text annotations using natural language techniques, or by using latent semantic analysis to match each word and surrounding words in the annotations to the different senses of the word provided by WordNet, among others.

More advanced techniques could have been used to generate more suited visual representations of the concepts. Some ideas are selecting more than one feature representation for each concept using Kohonen feature map on the feature vectors of the images associated to the concept [7], latent semantic analysis techniques applied to feature vectors as in [22], assigning weights to

the representations of concepts, and segmenting and extracting features of images at the region level instead of at the image level [13].

4.2. Use of the MediaNet Base

The extended content-based retrieval system is a typical content-based retrieval system that also includes a MediaNet knowledge base with media representations of concepts and a query processor that uses the MediaNet knowledge base to refine, extend, and translate user queries. Although the CB search engine only provides content-based retrieval based on color and texture features, the retrieval engine can accept text queries because MediaNet can be used to translate them into visual data. This is an added functionality to the retrieval engine provide by MediaNet. In this section, we shall describe how MediaNet is used by the query processor module to process queries in the different media integrated into the MediaNet knowledge base, in our case, visual and text queries.

When the user submits a query, either textual or visual, the query processor identifies relevant concepts to the query together with importance weights. This is done by matching the query to the media representations of the concepts in the knowledge base (e.g., a text query is matched to text representations of the concepts) and obtaining a relevance score for each concept indicating the similarity of the media representations of the concept to the query. The top ranked concepts and semantically similar concepts to these are considered relevant for the query. Then, the query and the media representations of the relevant concepts are matched to the image database by sending multiple queries to the CB search engine. Finally, the results are collected and merged into a unique list to display them to the user. The procedure is shown in Figure 1.

For image queries, color and texture feature vector are extracted from the query image and matched against the feature centroid of each concept using Euclidean distance. Only the percentage P% of the top ranked concepts are considered relevant to the query and selected. The feature vectors for all the images associated to each relevant concept are then matched to the query. Again, the percentage Q% of the top ranked images are considered relevant to the query and selected. The final dissimilarity score of each concept to the query is calculated as a function of the distances of the feature centroid of the concept to the query, the average distance of the relevant images of the concept to the query, and the proportion of the images of the concept that are relevant by the following equation:

$$fdist(q,c) = dist(q,cen_c) + \sum_{i \in rel_c} \frac{dist(q,i)}{num\_rel_c} * \sqrt{\frac{num_c}{num\_rel_c}} \quad (1)$$

where q is the feature vector of the query image, c is a relevant concept, cen_c is the feature centroid of concept c, num_c is the number of images of concept c, num_rel_c is the number of relevant images for concept c, rel_c are the feature vectors of the relevant images of concept c, i is a feature vector of one relevant image of concept c, and dist(q,c) is the weighted Euclidean distance among feature vectors q and c.

The top N most similar concepts to the query are then selected. The set of relevant concepts is expanded with at most M more concepts from the MediaNet knowledge base with lowest average conceptual dissimilarity to the relevant concepts. Only concepts with dissimilarities below R% the average conceptual distance in MediaNet can be selected as relevant. We only expand the set of relevant concepts with concepts that are similar enough semantically. For each concept added in this step, distance scores to the query are calculated as the sum of the average conceptual dissimilarity to the initial set of relevant concepts and the average distance of the initial set of relevant concepts to the query.

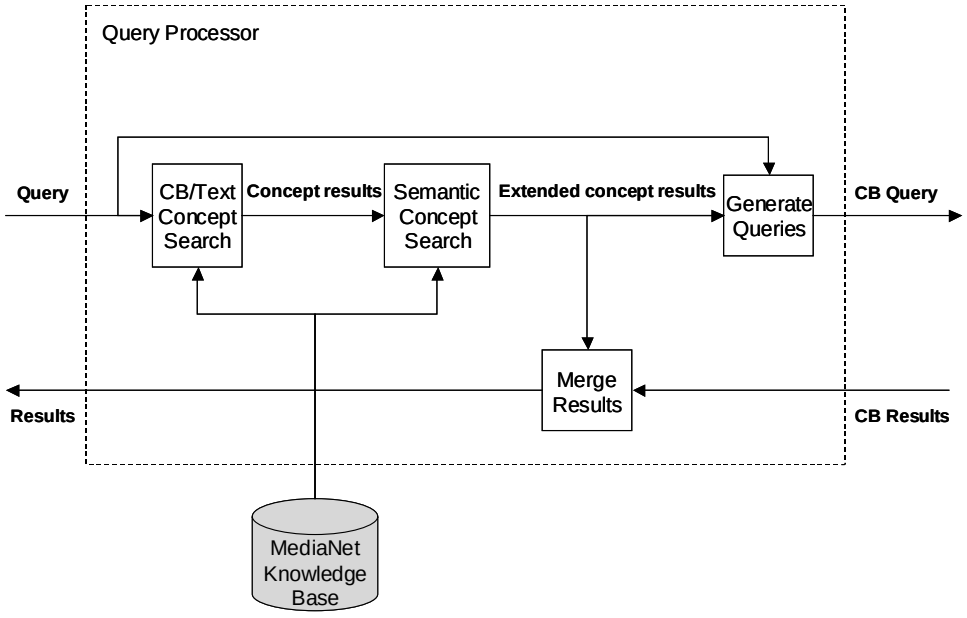

Figure 5: Procedure followed by the Query Processor.

The dissimilarity between two concepts is calculated based on the semantic relationships connecting the two concepts. Dissimilarity scores are assigned to the relationships connecting two concepts as follows. For antonymy, if two concepts are opposite, their dissimilarity is 1; if not, their dissimilarity is zero. For hypernymy/hyponymy, if two concepts are related through the Is-A-Subtype-Of/Is-A-Supertype-Of hierarchy, their dissimilarity is proportional to the number of concepts between the two in the hierarchy; if not, their dissimilarity is 1. For meronymy/holonymy, if two concepts are related through the Has-Part,Member,Substance/Is-Part,Member,Substance-Of network, their dissimilarity is proportional to the number of concepts between the two in the hierarchy; if not, their dissimilarity is 1. The proportional factor for the meronymy/holonymy relationship was set to twice the value of the factor for the hypernymy/hyponymy relationship. In other words, we are giving half the importance to the former relationship compared to the latter one. The total dissimilarity between two concepts is the average dissimilarity for the three relationships. Using this approach, the query can be both expanded and/or refined at the semantic level depending if the new concepts are broader (Is-A-Supertype-Of) and/or narrower (Is-A-Subtype-Of), respectively.

At this point, the query processor has a list of relevant concepts with dissimilarity scores to the query. It sends one query request for the image query and each relevant concept to the CB search engine. The feature vector of the query image is the input of the first query. The feature centroid of the relevant concepts is used as input for the other queries. The query processor merges the results of the multiple queries into a unique list of images and scores by summing the scores of each image in the database for the queries. The scores of the results of a concept are weighted in the summation by the following factor:

$$w(q,c) = 10^{-n * fdist(q,c)} \tag{3}$$

where q is the feature vector of the query, c is the feature vector of the concept, fdist(q,c) the final distance score of the query q and the concept c, and n is an positive real number. The larger the factor n is, the more importance is given to the concept representations to generate the final list of results.

For text queries, keywords are match to the list of synonyms of each concept. The feature centroid of the top concept is then treated as an input visual query and the relevant concept. Then, more relevant concepts are found based on conceptual dissimilarity, the queries are submitted to the CB search engine, and the results integrated into a unique list as described above for visual queries.

There is room for improvement in the procedure described above. Multiple relationship paths can connect two concepts could be considered. Then, the conceptual dissimilarity between two concepts could be calculated as the minimum weight of the paths. Single-value decomposition techniques such as latent semantic analysis could be used to obtain distance values between the query, and the database images and/or the concepts in the MediaNet base. Another new functionality supported a MediaNet knowledge base is browsing and navigation at the concept level. MediaNet also supports advanced inference, problem solving, and decision making tasks that are not implemented in the current retrieval system.

4. EXPERIMENTS

The MediaNet framework has been evaluated in an image retrieval application. The experiments setup and the experimental results are described in the following sections.

4.1. Experiments Setup

The objective of these experiments was to evaluate MediaNet in an image retrieval application. We compared the performance of the extended content-based retrieval engine that uses MediaNet to a base retrieval engine that does not use MediaNet. Both retrieval engines used the same image database and CB search engine; however, the MediaNet knowledge base and the query processor were only used in the former (see Figure 4). The image retrieval effectiveness [15] was measured in terms of precision and recall. 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 proposed to the standard [23]. 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 queries was also generated by MPEG-7. For each query, the ground truth represents a semantic, visual class and is annotated by a short textual description (e.g. “Flower Garden” and “News Anchor”). Initially, we used the ground truth generated by MPEG-7 in our experiments to compare the retrieval effectives of both systems but found it not suited. We, then, generated the ground truth including relevance scores for one query.

We used the textual descriptions associated with the ground truth of the queries to construct a MediaNet knowledge base as described in section 4.1. The total number of concepts derived from these textual annotations was 96. 50 of those concepts were related to other concepts by generalization/specialization (hypernymy/hyponymy) relationships; 34 concepts were related to other concepts by membership, composition, or substance (meronymy/hyponymy) relationships. There was only one case of antonymy. There were 387 images in the ground truth for all the queries. We generated the image and feature representations of the concepts using half of those images, which were not included in the image database.

Different experiments were performed with visual queries and text queries as input for different features. The values P, Q, and R were set to 20%; the values M and N to 3; and the value of n to 10 (see section 4.2). For the visual queries, we used the images specified by MPEG-7. We selected on word from each text annotation as the input to the text queries. Recall and precision results of the experiments are reported in the next section.

4.2. Experimental results

Figure 6 shows the precision and recall for the content-based retrieval and the extended content-based retrieval systems for the 50 queries and the ground truth generated by MPEG-7. For the extended content-based retrieval system, results for text and image queries are provided. In these experiments, both retrieval systems used all for features extracted from the images in the database with equal weights for retrieval. All the features were normalized to a maximum distance of 1 in the database.

As expected, the retrieval effectiveness for text queries in the extended content-based retrieval engine is much lower than for image queries. The results show a marginal gain of retrieval effectiveness for image queries by the extended content-based retrieval system. If two images are retrieved, the recall and the precision values for both systems are 0.5 and 0.8, respectively. These values are very high and are due to the fact that the ground truths contain images that are very similar visually (contiguous frames from the same video sequence, in most cases) and can be easily retrieved using low-level features. We run the same experiments using only color histogram and obtained similar results.

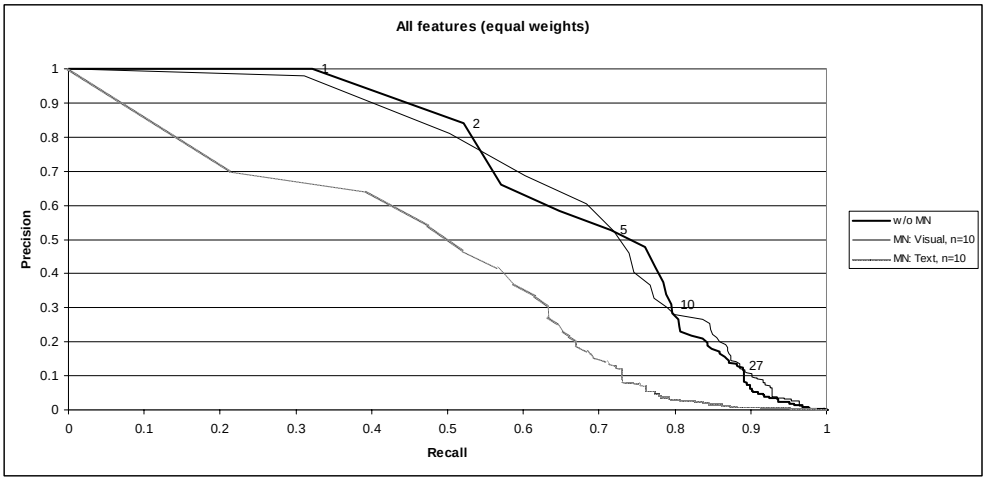

Figure 6: Average precision and recall for 50 queries. "w/o MN" corresponds to the content-based retrieval system that does not use MediaNet; "MN: Visual, n=10" to the extended content-based retrieval system with image queries; "MN: Text, n=10" to the extended content-based retrieval system using the keywords as queries. The numbers on the plot indicate the number of retrieved images for those values of precision/recall.

Figure 6 questions the suitability of the ground truth generated by MPEG-7 to evaluate the performance of the extended content-based retrieval. For this reason, we selected the query labeled as "Tapirs" and generated a more semantically meaningful ground truth. For the ground truth, relevance scores were assigned to all the images in the database as follows: "1" for pictures 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. The recall/precision values for both systems are shown in Figure 7. The extended content-based retrieval system provides better results than the content-based retrieval system for recall within the range (0.1, 1.7). These results are encouraging but additional experiments are needed.

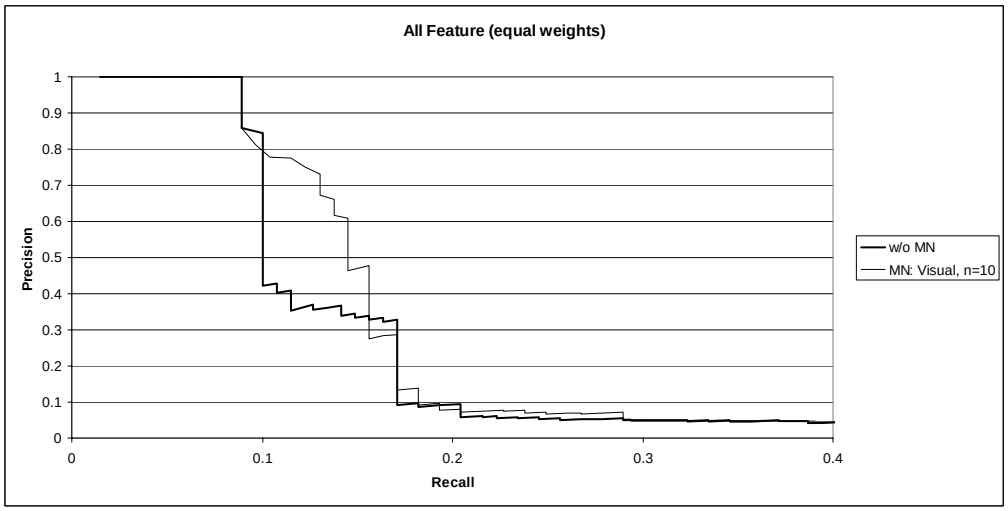

Figure 7: Average precision and recall for "tapirs" query and ground truth with relevance scores. "w/o MN" correspond to the content-based retrieval system that does not use MediaNet. "MN: Visual, n=10" corresponds to the extended content-based retrieval system using the images as queries.

5. CONCLUSIONS

We have presented MediaNet, which is a knowledge representation framework that uses multimedia content for representing semantic and perceptual information. The main components of MediaNet include conceptual entities and semantic and perceptual relationships among concepts, which are defined or exemplified by multimedia content. 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. We have reported on experiments that use MediaNet in searching for images. Initial experimental results demonstrate improved retrieval effectiveness using MediaNet in a content-based retrieval system.

REFERENCES

1. 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
2. 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.
3. M. Dobie, R. Tansley, D. Joyce, M. Weal, P. Lewis, and W. Hall, "A Flexible Architecture for Content and Concept Based Multimedia Information Exploration", *Proc. Of the Challenge of Image Retrieval*, pp. 1-12, Newcastle, Feb. 1999.
4. M. Flickner et al., "Query by Image and Video Content: The QBIC System", *Compute*r, Vol. 28, No. 9, pp. 23-32, Sep. 19995; also available at http://wwwqbic.almaden.ibm.com/.
5. D. W. Joyce, P. H. Lewis, R. H. Tansley, M. R. Dobie, and W. Hall, "Semiotics and Agents for Integrating and Navigating Through Media Representations of Concepts", *Proc. of Conference on Storage and Retrieval for Media Databases 2000, (IS&T/SPIE-2000)*, Vol. 3972, pp.120-31, San Jose, CA, Jan. 2000.
6. D. Lenat, "The Dimensions of Context Space", http://www.cyc.com/context-space.doc, Oct. 1998.
7. 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.
8. A. Meystel, *Semiotic Modeling and Situation Analysis: An Introduction*, AdRem, Bala Cynwyd, PA, 1995.
9. G. A. Miller, "WordNet: A Lexical Database for English", *Communication of the ACM*, Vol. 38, No. 11, pp. 39-41, Nov. 1995.
10. MPEG Multimedia Description Scheme Group, "MPEG-7 Multimedia Description Schemes XM (v4.0)", ISO/IEC JTC1/SC29/WG11 MPEG00/N3465, Beijing, CN, July 1999.
11. MPEG Multimedia Description Scheme Group, "MPEG-7 Multimedia Description Schemes WD (v4.0)", ISO/IEC JTC1/SC29/WG11 MPEG00/N3465, Beijing, CN, July 1999.
12. MPEG Requirements, "MPEG-7 requirements and objectives, whatever", XXX.
13. A. Natsev, A. Chadha, B. Soetarman, and J. S. Vitter, "CAMEL: Concept Annotated iMagE Libraries", Submitted.
14. M. R. Quillian, "Semantic Memory", Semantic Information Processing, M. Minsky (ed), MIT Press, Cambridge, MA, 1968.
15. J. R. Smith, "Quantitative Assessment of Image Retrieval Effectiveness", To appear in *Journal of Information Access*.
16. J. R. Smith and A. B. Benitez, "Conceptual Modeling of Audio-Visual Content", *Proc. Intl. Conf. On Multimedia and Expo (ICME-2000)*, July 2000.
17. J. R. Smith and S.-F. Chang, "VisualSEEk: A Fully Automated Content-Based Image Query System", *Proceeding of the ACM Conf. Multimedi*a, ACM Press, New York, 1996; also available at http://www.ctr.columbia.edu/VisualSEEk/.
18. S. W. Smoliar, J. D. Baker, T. Nakayama, and L. Wilcox, "Multimedia Search: An Authoring Perspective", *Proc. of the First International Workshop on Image Databases and Multimedia Search (IAPR-1996)*, pp. 1-8, Amsterdam, The Netherlands, Aug. 1996.
19. R. Tansley, "The Multimedia Thesaurus: An Aid for Multimedia Information Retrieval and Navigation", Master Thesis, Computer Science, University of Southampton, UK, Dec. 1998.
20. R. Tansley, C. Bird, W. Hall, P. Lewis, and M. Weal, "Automatic the Linking of Content and Concept", *Proc. of the conference on ACM Multimedia 2000*, Los Angeles, CA, Oct. 30 – Nov. 4, 2000.
21. W3C, "Extensible Markup Language (XML)", http://www.w3.org/XML/.
22. R. Zhao, and W. I. Grosky, "From Features to Semantics: Some Preliminary Results", *Proc. of IEEE International Conference on Multimedia and Expo 2000*, New York, NY, July 30 – Aug. 2, 2000.
23. D. Zier, J.-R. Ohm, "Common Datasets and Queries in MPEG-7 Color Core Experiments", ISO/IEC JTC1/SC29/WG11 MPEG99/M5060, Melbourne, Australia, Oct. 1999.