Multimedia Knowledge:

Discovery, Classification, Browsing,

and Retrieval

Ana Belén Benítez Jiménez

Submitted in partial fulfillment of the

Requirements for the degree

of Doctor of Philosophy

in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2005

ABSTRACT

Multimedia Knowledge:

Discovery, Classification, Browsing, and Retrieval

Ana Belén Benítez Jiménez

Humans behave and reason by learning and maintaining dynamic models of the world. In addition, there is evidence from psychology that human models contain textual nodes (semantic; word "watermelon") and audio-visual nodes (perceptual; quality of being green). This is why in an attempt to make sense of multimedia as humans do, this thesis focuses on representing and discovering semantic and perceptual knowledge about the world using multimedia and knowledge from external resources such as the electronic dictionary WordNet. This thesis also proposes techniques that use the discovered knowledge to advance multimedia classification, browsing, and retrieval applications.

In this thesis, we propose a unified framework, MediaNet, that uses multimedia for representing semantic and perceptual knowledge in the form of concepts networks with media examples, which we call "medianets". The MediaNet framework extends semantic knowledge frameworks such as thesaurus and ontologies by including perceptual knowledge, and exemplifying concepts and relationships using multimedia. We use the MPEG-7 standard to represent medianets in an interoperable way.

This thesis also proposes new techniques for discovering and summarizing medianets from annotated images. Medianets are constructed by clustering images based on visual and textual features; and disambiguating the senses of words in annotations using WordNet and the image clusters. Visual, statistical, and semantic relationships are then discovered between clusters and senses (i.e., concepts). Finally, medianets can be summarized by merging statistically similar concepts. In contrast to prior work, we integrate the processing of images and annotations to improve the concept discovery. In addition, our summarization techniques are general, automatic, and applicable to any domain.

Image classifiers can be used to annotate images with semantic labels such as "person", "mountain", and "outdoors". A key problem of current image classification systems is the lack of systematic methods for finding relevant classes and their relations. We propose the use of extracted medianets for automatically discovering salient classes and combining several individual detectors for superior accuracy (up to a 15% gain). We train individual detectors to predict the presence of concepts in images, which are combined statistically using a Bayesian network based on the medianets.

Image browsers enable users to gain a quick insight into the content of a collection by supporting several exploration tasks (e.g., locating images). Current approaches organize images on large and complex structures, which result in inefficient navigation. Our approach organizes annotated images in a multi-resolution hierarchy of medianets obtained by iteratively clustering similar concepts in the extracted medianet. Users can

browse the collection by navigating multimedia visualizations of the resulting hierarchies. We have demonstrated our browsing system is significantly more useful, easier to use, more stimulating, and more successful, among others, in an extensive user study.

Extracted medianets also enable new ways of searching for images using multiple modalities. In particular, we use medianets for expanding, refining, and translating user queries across different modalities in an image retrieval system. First, we detect relevant concepts in incoming queries and add other semantically and perceptually similar concepts. Then, images are retrieved and ordered based on how closely they match the incoming query and the relevant concepts. Initial experiments have demonstrated improved retrieval effectiveness using the proposed techniques in a semantic query.

Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, I would like to thank my advisor Prof. Shih-Fu Chang. To him, I owe the awakening of my interest on multimedia information systems. I still remember the thrill of working on the course projects from his Visual Information System class. He has coached me during the long and hard path of my doctorate and I really appreciate it.

Second, I would like to extend my gratitude to Dr. John R. Smith who has been a constant inspiration and support from the beginning of my doctorate. The topic of my thesis was born during an internship in his group at IBM T. J. Watson Research Center. In addition, we worked together in numerous MPEG meetings.

I would also like to thank the rest of the members in my defense committee. Prof. Alexandros Eleftheriadis was actually my first advisor in my doctorate. Dr. Nevenka Dimitrova offered me an internship at Philips at the very beginning without knowing me. Finally, Prof. Ellis help shaped my thesis by participating in my defense proposal.

Financially, my doctorate was supported by "la Caixa" and Kodak fellowships. In this regard, I need to thank Dr. Alex Loui for being my mentor during my time as a Kodak fellow. In addition, I greatly appreciate Dr. Chung-Sheng Li and Yoshihisa Gonno for the chance of working by their side as an intern at IBM and Sony, respectively.

Now, I would like to thank friends and colleagues with whom I had the pleasure to work during my doctorate. Many were the nights I spent working at Columbia University with Dr. Alejandro Jaimes (on the way to becoming a proud parent now). Also, nights were always full of interesting conversations with Prof. Hari Sundaram and Raj Kumar.

Other people whom I worked with, or alongside at, Columbia University were Javier Zamora, Mandis Beigi, Seungyup Paek, Di Zhong, Ryoma Oami, Qibin Sun, Ching-Yung Lin, William Chen, Shahram Ebadollahi, Winston Hsu, Yong Wang, Lexing Xie, Dongqing Zhang, Tian-Tsong Ng, and Paul Bocheck, among others.

In MPEG, I was an honor to meet and work with people such as Prof. Philippe Salembier, Toby Walker, Dr. Hawley Rising III, Prof. Corinne Jörgensen, Prof. José M. Martinez, Dr. Atul Puri, Dr. Qian Huang, Dr. Charlie Judice, Dr. Yoshiaki Shibata, Dr. Masahiro Shibata, Jörg Heuer, Dr. Jane Hunter, Dr. Ajay Divakaran, among many others.

Last but not least, I would like to thank my friends and family for their love, attention, and continuous inquiries about my graduation. Such a short paper for such a long list: Olga Enriquez, Geoffrey K. Mize, Dr. Javier Gomez, Kodyr Kolmatov, Dr. Rahul Swaminathan, Dr. Ralf Mekle, Dr. Masako Mori, Dr. Deepa Majmudar, and many others.

Today, on Saint Valentine's Day, I am very fortunate to have Arvid J. Bessen at my side. He has shared with me a very eventful and exciting transition towards graduation. I can only hope he keeps bearing with me for a long time to come.

Mamá, Papá, Javi, Inma, y Mari sé que no os lo vais a creer pero finalmente terminé el doctorado. ;-)

To my family and friends

1 Introduction

1.1 Overview

We humans experience the world through our senses. We wake up in the morning to the contact of soft sheets, the smell of freshly brewed coffee, and the noise of neighbors upstairs. As the day progresses, we undertake the most common tasks in life by analyzing and interpreting data we capture using multiple senses. For example, we pick a watermelon in a grocery store based on looks, smell, touch, and sound. This is possible only because we learn and maintain dynamic models of our knowledge of the world that allow us to understand, behave, and reason in almost any situation. In particular, I learned to pick up watermelons looking at my mother in the town market. She always managed to find the best ones!

Multimedia such as images, video, and audio are equivalent to human sensory data. The sensors, in this case, are recording devices such as photographic cameras, video camcorders, and audio recorders. Understanding and reasoning on sensory data are innate skills in humans, who use them in their everyday life. However, these are extremely difficult tasks for multimedia information systems that, in an analogous way, aim at managing multimedia. The difficulty is predicated on many open issues in artificial

intelligence, semiotics, natural language processing, computer vision, and auditory analysis, among others.

The main goal of this thesis is to improve current multimedia information systems by attempting to represent, analyze, and utilize multimedia as humans do with sensory data: by building knowledge models of the world, and by understanding, behaving, and reasoning using those models. In particular, this thesis proposes to represent and discover knowledge about the world using multiple media and knowledge from external resources. It also investigates ways of using this multimedia knowledge to improve classification, browsing, and retrieval applications of multimedia. Knowledge is defined as facts about the world and represented as nodes and arcs between the nodes, i.e., semantic networks (e.g., watermelons are a fruit and that they are green and round; see Figure 1.1).

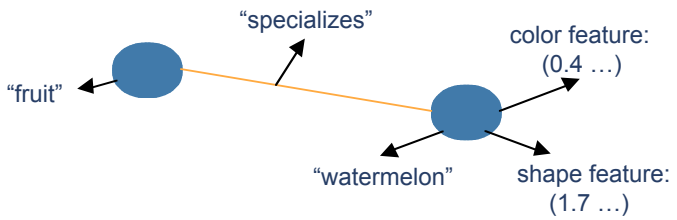


Figure 1.1: Example of a semantic network that represents the statements "watermelons are a fruit and they are green and round". The semantic network contains two concepts "fruit" and "watermelon" (circles) and a semantic relationship "specializes", which relates the two concepts (non-arrow line). The concepts and the relationship have word examples (e.g., "fruit" and "watermelon"). In addition, the concept "watermelon" has color and shape features representing its green and round attributes, respectively.

The specific problems addressed by this thesis follow. We focus on multimedia in the form of images with annotations.

- **Representation**: The goal of this problem is to represent perceptual and semantic knowledge (i.e., information about what our senses perceive and the meaning of such perceptions, respectively) about the world using multimedia. We encode the multimedia knowledge in a flexible, extensible, and interoperable way.

- **Discovery**: In this problem, we extract perceptual and semantic knowledge from a collection of annotated images. In this process, we integrate the processing of both images and annotations, and reuse knowledge from external resources. We also automatically summarize and evaluate the extracted multimedia knowledge.

- **Classification**: This problem aims at annotating new images with semantic labels or classes (e.g., "vegetation"). We use the knowledge extracted in the discovery process for automatically discovering relevant classes and combining classifiers to improve the classification accuracy.

- **Browsing**: This problem tackles the organization and display of an image collection for browsing. We organize the images based on multi-resolution summaries of the extracted knowledge. Both images and words are used to display the hierarchy of knowledge summaries.

- **Retrieval**: In this problem, we retrieve relevant images from a collection in answer to multimedia queries from users. We use extracted knowledge to expand, refine, and/or translate incoming queries to other modalities, as needed.

This thesis proposes solutions to the problems listed above that invariably integrate perception and semantics. For example, we integrate the processing of images and words, and combine perceptual and semantic knowledge in the same framework, among others. The intuition comes from human mental models, which seem to contain nodes of not only textual (semantic) but also audio-visual nature (perceptual). As an example (see Figure 1.1), humans know that watermelons are a fruit (semantic) and that they are green and round (perceptual). The proposed techniques are also automatic, generic, and applicable to different types of media (e.g., text and images) and knowledge (e.g., perceptual and semantic).

All the methods presented in this thesis have been developed and used within the IMKA (Intelligent Multimedia Knowledge Application) framework. The IMKA framework is a novel and flexible system for developing and evaluating multimedia information systems. We evaluated the proposed techniques using standard evaluation measures such as classification accuracy and retrieval effectiveness, among others. In addition, we conducted an extensive user study for evaluating the efficiency, the effectiveness, and the subjective satisfaction of the proposed knowledge-based browsing techniques. We used diverse collections of annotated images from different domains including scientific (nature), news, and consumer (travel) images.

1.1.1 Outline of the Chapter

The rest of the chapter is organized as follows. In the next section, we make a case for the need of knowledge-based approaches in multimedia information systems. In section 1.3, we summarize the intuition behind, and the solution to the problems mentioned in the previous section. In section 1.4, we outline the contributions of this thesis. Finally, section 1.5 concludes with an outline of the thesis.

1.2 Motivation

In recent years there has been a substantial proliferation of available multimedia for personal, entertainment, and scientific uses. Much of this multimedia is in digital form due to the increasingly cheap and accessible technologies for recoding, storing, accessing, and exchanging this kind of data. The ubiquity of multimedia requires effective multimedia information systems that help to manage the large amounts of multimedia.

The most effective current information systems rely almost exclusively on text in spite of the many open problems remaining in natural language processing. Who does not know Google? In comparison, the capabilities of emerging audio- and visual-based systems are very limited due to the added complexity of the audio and image processing. Recent research on the analysis of audio-visual data has enabled multimedia information systems that support Content-Based Retrieval (CBR) [3][48][135][121][131][155], automatic but constrained classification of objects and scenes [49][104][105][106][110] [111][143][150][154][164], and simple mechanisms for learning relevance feedback from

users [122][172] and for labeling images and regions [5][47][100][106]. 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.

Numerous content-based retrieval systems have explored the possibilities of indexing images and videos by using low-level visual features [121][131][155] (for example, see Virage [3], QBIC [48], and VisualSEEk [135]). These systems work by (1) automatically extracting features directly from the visual data; (2) indexing the extracted features for fast access; and (3) querying and matching features for the retrieval of the visual data. Beyond these basic capabilities, there have been efforts to support relevance feedback from users to refine queries and to learn through positive and/or negative examples what the user might be looking for [122][172].

As users are mostly interested in retrieving images at the semantic level (e.g., people and location of images) [1][73], more recently, there has been an effort on automatically producing certain semantic labels that could contribute significantly to retrieving visual data. For example, there has been work on detecting indoor vs. outdoor, city vs. landscape, faces and people, etc [49][110][111][143][154][164]; and other attempts to assign categories from larger controlled vocabularies about the who, what, when, and where of visual data (e.g., people, horse, house, mountain, and town) [28][104][105][150][151]. More recent methods aim at extending image annotations to new images or regions [5][47][100][106]. Most of these approaches rely on the

application of machine learning techniques to computer vision. These systems usually retrieve images using the class vectors (1 if label is present, 0 otherwise) or textual features extracted from the image annotations sometimes in conjunction with extracted visual features [54][137][138]. High degree of success has been reached for various domain-specific and constrained test sets.

This thesis takes a different approach towards improving multimedia information systems. We believe that advanced multimedia information systems also need to be capable of communicating with the user and of understanding multimedia 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. However, for any communication to take place between two parties, there must be a common context or knowledge shared between the parties (e.g., watermelons are a fruit and that they are green and round). Humans interact, understand, behave, and reason by learning and maintaining dynamic models of their knowledge of the world. In addition, there is evidence from psychology that human mental models contain nodes of not only textual (semantic) but also audio-visual (perceptual) nature [44][70][79][123]. That is why we propose a novel multimedia information system that is completely driven by perceptual and semantic knowledge extracted from multimedia.

Knowledge is usually defined as facts about the world and represented as nodes and arcs between the nodes, i.e., semantic networks (see Figure 1.1). Nodes represent objects,

concepts, events, or situations (e.g., watermelon and fruit), and the links represent relationships between nodes (e.g., watermelon is a fruit). Knowledge can refer to either perceptions (how our senses perceive things, e.g., some green pattern) or semantics (how we interpret or assign meaning of things, e.g., meaning of word "watermelon"). Therefore, we distinguish two different kinds of knowledge, perceptual and semantic knowledge. We believe that perceptual and semantic knowledge need to be unified to impact multimedia information systems, which have to deal with multimedia at the semantic and perceptual levels, as humans do.

The approaches proposed in this thesis are the first of their kind to represent and automatically discover generic perceptual and semantic knowledge about the world using multimedia and knowledge from external resources; and to classify, browse, and retrieve multimedia automatically exploiting the extracted knowledge. The novel multimedia knowledge representation framework is presented in chapter 3. We discover multimedia knowledge from annotated images by integrating the processing of images and annotations, and by reusing knowledge from the electronic dictionary WordNet, as we present in chapter 4. We have found that our knowledge-based approach can improve the performance of image classification, browsing, and retrieval. In chapters 5, 6, and 7, we report on experiments that demonstrate the advantage of using multimedia knowledge in classifying, browsing, and searching images, respectively.

We show that these techniques enable a step towards the analysis of media beyond text and the development of fully integrated, multi-modality systems. These systems would

be able to enrich human experiences revolutionizing a variety of application scenarios dealing with education, commerce, and entertainment. When Google allows retrieving web pages using queries that combine text, images, and audio, and displaying the results in a multimedia form, humans will start truly learning from, and interacting with the world beyond the barriers of physical surroundings and individual experiences.

1.3 Problems Addressed and Proposed Approaches

In this section, we present the intuition behind our solutions to the problems of multimedia knowledge representation and discovery, and of image classification, browsing, and retrieval using the extracted knowledge. All the methods presented in this thesis have been developed and used within the IMKA (Intelligent Multimedia Knowledge Application) framework. The IMKA framework also supports the design, configuration, test, and evaluation of generic multimedia information systems.

1.3.1 Multimedia Knowledge Representation

In this thesis, we propose a novel framework for representing knowledge using multimedia, MediaNet. Existing frameworks lack the expressive power to represent the rich knowledge in human mental models.

MediaNet is an integrated knowledge representation framework, the first of its kind to use multimedia for representing generic perceptual and semantic knowledge about the world [22][23]. We define perceptual and semantic knowledge as follows; which are

adapted from the definitions of percept vs. concepts and syntax vs. semantics in conceptual framework [64].

- **Perceptual Knowledge**: This knowledge refers to what our senses perceive. For example, we perceive light through our eyes. Patterns of light produce sensations such as texture and color without the need for interpreting the patterns. Therefore, in terms of images, this knowledge can be represented using visual features extracted from the images.

- **Semantic Knowledge**: This knowledge refers to the interpretation or meaning of what is perceived. For example, a specific color and texture pattern can be identified as being a specific object such as a fruit. Therefore, semantic knowledge is best represented using words in a language.

The main components of the MediaNet framework include (1) concepts, (2) relationships between concepts, and (3) media exemplifying concepts and relationships such as images, words, and low-level features of the media. In other words, the MediaNet framework represents knowledge as concept networks with media examples,

which we refer to as "multimedia concepts networks" or "medianets"[i]. Figure 1.2 shows an example of a medianet exemplifying concepts "fruit" and "watermelon" illustrated by text, an image, and a color feature. Semantic relationship "specializes" relates the two concepts. The MediaNet framework extends and differs from related work in two ways: (1) in combining concepts and relationships at the perceptual and semantic levels in the same network, and (2) in illustrating perceptual and semantic relationships using media.

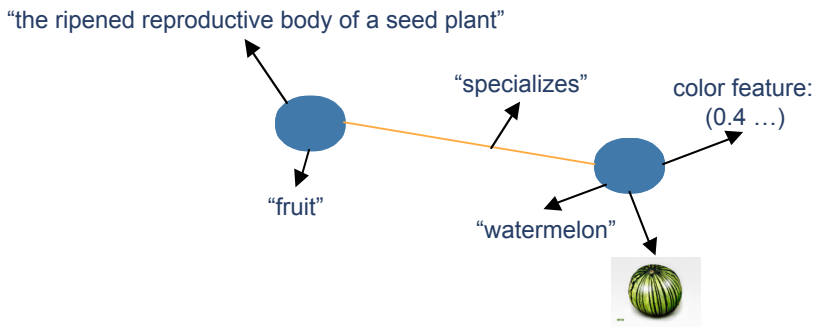


Figure 1.2: Example of a medianet that represents concepts "fruit" and "watermelon" (circles) illustrated by text, an image, and a color feature (arrow lines). Semantic relationship "specializes" relates the two concepts (non-arrow line).

In designing the MediaNet framework, we have built on the basic principles of semiotics [93] and semantic networks [116], in addition to utilizing evidence from psychology that

[i] Please, note that we use the capitalized word "MediaNet" and "MediaNet framework" to refer to the proposed multimedia knowledge representation framework; whereas, the non-capitalized word "medianet" is used to mean "multimedia concept network".

humans have rich mental models of the world that contain interconnected nodes of textual and audio-visual nature [44][70][79][123]. The MediaNet framework offers functionality similar to that of a dictionary, ontology, or thesaurus by defining and describing semantic concepts. However, it extends these frameworks by denoting concepts and relationships between concepts at the semantic and perceptual levels using multimedia. We use the MPEG-7 standard [61][87] for multimedia description to encode, represent, and exchange medianets in a machine processable, reusable, and interoperable way. Medianets are encoded using MPEG-7 tools that describe structured collections [133] (e.g., collections of images), the structure [19] (e.g., regions in images), and the semantics [19] (e.g., people depicted in an image) of multimedia.

1.3.2 Multimedia Knowledge Discovery

This thesis also focuses on the discovery, summarization, and evaluation of multimedia knowledge from annotated images such as image clusters, word senses, and relationships among them. Existing techniques process each media independently or are domain specific so they do not generalize to arbitrary types of media or knowledge.

The substantial proliferation of multimedia such as annotated images requires tools for discovering relevant knowledge (i.e., concepts and relationships) from annotated images to enable innovative and intelligent classification, browsing, and retrieval of images. Perceptual knowledge (e.g., image clusters and relationships between them) is essential because it can be extracted automatically and is not inherently limited to a set of words

as textual annotations. On the other hand, semantic knowledge (e.g., word senses and relationships between them) is the most powerful because human communication often happens at this level. However, current approaches for extracting semantic knowledge are, at best, semi-automatic and very time consuming. Furthermore, the discovered knowledge may be too large and unstructured to be useful; therefore, it is often necessary to summarize multimedia knowledge in order to reduce it size while maintaining the essential knowledge. Hence, automatic ways to quantify the consistency, the completeness, and the conciseness of the multimedia knowledge, among others, are essential to evaluate and compare different techniques for knowledge discovery and summarization. As an example, knowledge is consistent if contradictory conclusions cannot be reached from valid input definitions.

This thesis proposes novel and automatic methods for discovering perceptual and semantic knowledge from annotated images, and for summarizing, and evaluating arbitrary multimedia knowledge [10][14][15][16]. In the multimedia knowledge discovery process, a medianet is built for a collection of annotated images. We extract perceptual knowledge [10][15] from a collection of annotated images by discovering perceptual concepts through clustering the images based on visual and textual features. We find perceptual relationships among the clusters based on feature similarity and statistical dependencies between the clusters. In the semantic knowledge extraction process [10][16], we discover semantic concepts by disambiguating the sense of words in the annotations using WordNet [95] and the image clusters. We find semantic relationships

among the detected senses based on WordNet. We facilitate and automate and the knowledge discovery process by integrating the processing of images and annotations, and by reusing knowledge from exiting resources such as WordNet.

The proposed knowledge summarization techniques reduce the size of a medianet (in terms of number of concepts and relationships) by grouping similar concepts together [10][14]. We calculate the distance or dissimilarity between concepts using a novel concept distance measure based on concept statistics and network topology (e.g., concepts that are adjacent in the network and share large conditional probabilities have small distances). We also propose automatic techniques for measuring the consistency, the completeness, and the conciseness of medianets based on notions from information and graph theory such as entropy and graph density [10][14]. As an example, we propose to calculate the consistency of a medianet as the spread of the distances between concepts through different paths in the medianet.

We have performed several experiments for evaluating the proposed techniques for perceptual and semantic knowledge discovery, and for multimedia knowledge summarization. Experiments have shown the superiority of integrating the analysis of images and annotations for improving the performance of the image clustering and the word-sense disambiguation, the importance of good concept distance measures for knowledge summarization, and the potential of automatic measures for evaluating knowledge quality.

1.3.3 Knowledge-Based Multimedia Classification, Browsing, and Retrieval

The problems of classifying, navigating, and searching multimedia are fundamental in the multimedia literature, and their solution helps in managing and accessing large collections of multimedia. Most of the existing techniques rely solely on features directly extracted from the multimedia. Thus these approaches are limited by the capabilities of current media analysis and processing tools. Our approach has been to explore novel ways of exploiting automatically extracted medianets from annotated images, including knowledge from external resources such as the electronic dictionary WordNet, to enhance the classification, browsing, and retrieval of images.

Image classifiers can be used to annotate images with semantic labels such as "person", "mountain", and "outdoors". Recent advances in image classification have explored the interaction among multiple classes. However, a key problem still remains the lack of systematic methods for finding relevant classes and their relations. Central to our image classification work is the use of extracted (and summarized) medianets for automatically discovering salient classes and combining several individual detectors [12]. Individual detectors are trained to predict the presence of concepts in images, which are combined statistically using a Bayesian network based on the medianets. Experiments have shown that statistically combining classifiers using extracted medianets can result in superior accuracy (up to a 15% gain) compared to individual classifiers or purely statistically learned classifier structures.

Image browsers enable users to gain a quick insight into the content of a collection by supporting a variety of exploration tasks (e.g., locating images related to a specific topic). Current approaches organize images based on low-level features (e.g., color features) or semantic structures (e.g., concept networks) that might be too large and complex for efficient navigation. Our approach for multimedia browsing makes use of the extracted medianets to organize annotated images in multi-resolution networks [13]. At the highest resolution, we organize the images using the medianet discovered from the annotated images. Medianets at lower resolutions are constructed by clustering similar concepts together. Users can browse the annotated images by navigating multimedia visualizations of the resulting hierarchies of medianets in a non-strictly hierarchical fashion. We conducted experiments with real users that demonstrated the superior effectiveness (e.g., better images selected for two travel pamphlets), efficiency (e.g., fewer viewed images and executed browsing operations), and subjective satisfaction (e.g., more useful, easier, more stimulating, and more successful) of users in performing common browsing tasks using the proposed techniques.

Extracted medianets also enable new ways of searching for images using multiple modalities. In particular, we use medianets for expanding, refining, and translating user queries across different modalities or media forms in an image retrieval system [22][23]. First, we classify incoming queries, either words or images, into relevant semantic and perceptual concepts using concept detectors. The initial set of detected concepts is extended with semantically and perceptually similar concepts. Finally, images are

retrieved that match the final set of concepts and their associated words and images. The retrieved images are ordered based on how closely they match the user query and the concept set. Higher importance is assigned to the initial set of detected concepts than the additional concepts. Initial experiments have demonstrated improved retrieval effectiveness using the proposed techniques in a semantic query.

1.4 Summary of Contributions

We now summarize the original contributions of this thesis when solving the problems of multimedia knowledge representation and discovery, and of classification, browsing, and retrieval of images using extracted knowledge.

1.4.1 Multimedia Knowledge Representation

1. A knowledge representation framework, MediaNet, which uses multimedia to represent knowledge about the world in the form of medianets, i.e., concept networks with media examples.

2. The idea to combine perceptual and semantic knowledge in the same concept network imitating, thus, the way humans build models of the world with perceptual and semantic information.

3. Encoding knowledge represented by the MediaNet framework using MPEG-7 structured collection, structure, and semantic description tools for reusability and interoperability by other applications.

1.4.2 Multimedia Knowledge Discovery

1. The idea to integrate the processing of images and annotations to discover knowledge about a collection of annotated images in form of medianets, i.e., concepts networks with media examples.

2. Finding novel similarity and statistical relationships between perceptual concepts or image clusters based on statistics and visual similarities.

3. Disambiguating the senses of words in image annotations using the electronic dictionary WordNet and the image clusters.

4. The idea of automatically summarizing knowledge by clustering similar concepts together.

5. A concept distance measure for arbitrary medianets based on the topology of the network and the statistics of concepts.

6. Automatic techniques for measuring the consistency, completeness, and conciseness of medianets based on notions from information and graph theory.

1.4.3 Knowledge-Based Multimedia Classification, Browsing, and Retrieval

1. The idea of using medianets for improving current classification, browsing, and retrieval applications for images.

2. Classification:

    a. The idea of discovering relevant classes (concepts) using extracted (and summarized) medianets for a collection of training images with annotations.

    b. The idea of building individual concept detectors using extracted (and summarized) medianets.

    c. Building a Bayesian network from medianets to combine the individual concept detectors.

    d. Using the Bayesian network of detectors to label new images with relevant concepts.

3. Browsing:

    a. The idea of organizing a collection of annotated images in a multi-resolution fashion using a hierarchy of medianets, i.e., extracted and summarized medianets.

    b. Displaying the multi-resolution hierarchy of medianets in a graphical way by combining images and words.

    c. The idea of enabling users to browse the medianet hierarchy in a non-strictly hierarchical fashion.

   d. An extensive user study to evaluate the effectiveness, efficiency, and subjective satisfaction of real users in performing common browsing tasks with the proposed techniques compared to related work.

4. Retrieval:

   a. The idea of using medianets to expand and refine queries of different modalities.

   b. The idea of using medianets to translate queries across different modalities.

   c. A system that retrieves images by processing the queries and the results using a medianet.

5. Implementing and testing all the proposed techniques in the IMKA framework.

6. A framework for designing, configuring, testing, and evaluating generic multimedia information systems, the IMKA framework.

1.5 Outline of The Thesis

The rest of this thesis is organized as follows. In the next chapter, we review relevant literature to the problems of knowledge representation, discovery, and application using multimedia. In addition to traditional frameworks for knowledge representation and

understanding, we present semantic, perceptual, and integrated frameworks for representing knowledge.

In chapter 3, we present our framework for representing knowledge about the world using multimedia, MediaNet. First, we discuss the most relevant work on using multimedia to represent knowledge, and also present evidence from psychology on multimedia mental models in humans. We conclude the chapter with a discussion of the mapping of the MediaNet framework to semantic networks, semiotics, and MPEG-7.

In chapter 4, we present our research on multimedia knowledge discovery. There, we present the specific algorithms for extracting semantic and perceptual knowledge from annotated images. We also describe our techniques for automatically summarizing and evaluating medianets. We conclude the chapter with a discussion of the experimental evaluation of these techniques.

In chapter 5, we present our approach towards knowledge-based image classification. We focus on three aspects related with the extracted medianets: (1) finding relevant classes and their interactions in a domain; (2) building a Bayesian network classifier for the classes; and (3) labeling new images using the resulting classifier. The chapter concludes with the experimental evaluation of the proposed image classification techniques.

Chapter 6 describes our work on knowledge-based image browsing. First, we describe the multi-resolution organization of image collections as medianet hierarchies. Then, we present the system that displays the medianet hierarchies and allows users to browse that

structure in a non-strictly hierarchical way. Finally, this chapter presents, in detail, the user study performed for evaluating the effectiveness, efficiency, and subjective satisfaction of users in performing common browsing tasks with the proposed system.

In chapter 7, we introduce the problem of retrieving images using extracted medianets. In addition to discussing some experimental evaluation, we present our approach to expand, refine, and translate queries from users across different modalities or media forms for image retrieval.

In chapter 8, we present the IMKA framework that implements the techniques proposed in this thesis including the software architecture and some sample systems. In chapter 9, we present the conclusions of our work as well as discuss some future research directions. Chapter 10 contains the papers and work referenced in this thesis. Finally, chapter 11 provides an overview of relevant MPEG-7 tools for this thesis and an example of the questionnaire used for evaluating the proposed knowledge-based browsing techniques.

2 Basic Concepts and Literature Review

2.1 Introduction

This chapter reviews relevant literature on knowledge representation, discovery, and application using multimedia. Knowledge representation consists on representing facts about objects, concepts, and events of the world, among others. In particular, we review traditional knowledge representation models and more recent works that use different media for knowledge representation. In addition, we discuss some proposed approaches for extracting knowledge from multimedia and for using that knowledge in multimedia applications such as image retrieval and browsing, among others.

Knowledge is usually defined as facts about the world and represented as nodes and arcs between the nodes, i.e., semantic networks. The nodes represent objects, concepts, events, or situations in the world (e.g., "watermelon" and "fruit"), whereas, the links represent relationships between nodes (e.g., watermelon "is a" fruit). Apart from semantic networks, traditional models for knowledge representation include logics, frames, and scripts. We also turn to semiotics to understand the creation and the role of knowledge in human intelligent behavior and thinking.

We categorize specific frameworks for knowledge representation into semantic, perceptual, and integrated frameworks. Semantic frameworks use text (i.e., words) to represent knowledge about the world such as classical thesauri. On the other hand, perceptual frameworks use audio and/or visual data (e.g., images and texture features) for knowledge representation; an example of perceptual framework is the Visual Thesaurus. Integrated frameworks such as the Multimedia Thesaurus represent knowledge using textual and audio-visual data. The structured collection description tools in the MPEG-7 standard [61][87] could also be considered an integrated knowledge representation framework. We provide an overview of these and other MPEG-7 description tools relevant for this thesis in appendix 11.1. We do not present it in this section because MPEG-7 is used as the meta-language to encode and represent knowledge in the proposed MediaNet framework.

The way knowledge is extracted differs in perceptual, semantic, and integrated frameworks. Semantic knowledge is often constructed manually by experts; whereas perceptual knowledge can be usually extracted automatically by the processing, segmentation, and clustering of audio and/or visual data. Finally, integrated knowledge is often discovered semi-automatically through interactions with users and experts. These frameworks have been used successfully in information retrieval and browsing; for example, to expand keyword queries in a web search engines using textual thesauri.

2.1.1 Outline of the Chapter

The rest of the chapter is organized as follows. In the next section, we present some traditional models for knowledge representation and understanding. In section 2.3, we review semantic frameworks for knowledge representation. In section 2.4, we describe perceptual frameworks. We present integrated frameworks in section 2.5. Finally, we conclude with a summary of the chapter in section 2.6.

2.2 Knowledge Representation and Understanding

Traditional knowledge representation models developed in the field of artificial intelligence are logics [124], semantic networks [116], frames [98], and scripts [126]. For insights into knowledge understanding, we turn to semiotics [93], the science that studies the understanding of signs such as words and pictures by humans.

2.2.1 Logics and Semantic Networks

Logics and semantic networks are widely accepted models for effective knowledge representation.

Logics [124] aim at emulating the laws of thought by providing a mechanism to represent statements about 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 assumptions

about 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) [124], also known as First-Order Predicate Logic (FOPL), which assumes (1) the existence of facts, objects (individual entities), and relationships among objects; and (2) the beliefs of true, false, and unknown for statements. For example, "Brother(Richard, John) Λ Brother(John, Richard)" means that "Richard is the brother of John and John is the bother of Richard"; "∀x King(x) ⇒ Person(x)" means that "All kings are persons".

Semantic networks [116] 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 cannot represent negation or disjunction, among others. Other knowledge representation models that have also been proven to be effective in knowledge representation are frames [98] and scripts [126], which distinguish between stereotypical and instance situations. A frame is a network of nodes and relations whose higher levels represent attributes that are always true about a situation and whose lower levels contain information about specific instances of the situation. 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. It is widely accepted that knowledge in the form of semantic networks, frames, and scripts can be expressed using logics such as First-Order Logic.

2.2.2 Semiotics

Semiotics [93] 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 logics. Our interest in semiotics lies in trying to understand the creation and the role of knowledge in human intelligent behavior and thinking. Several works on multimedia information systems have already exploited and incorporated ideas from semiotics.

For our interests, a more specific and 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" [93]. 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 2.1). In addition, the modeling of the unit of intelligence gives rise to multiresolution.

Figure 2.1 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 (i.e., using signs). 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 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 creates changes in the world, physically and/or conceptually. New objects emerge, which can be again encoded by sensors closing the cycle.

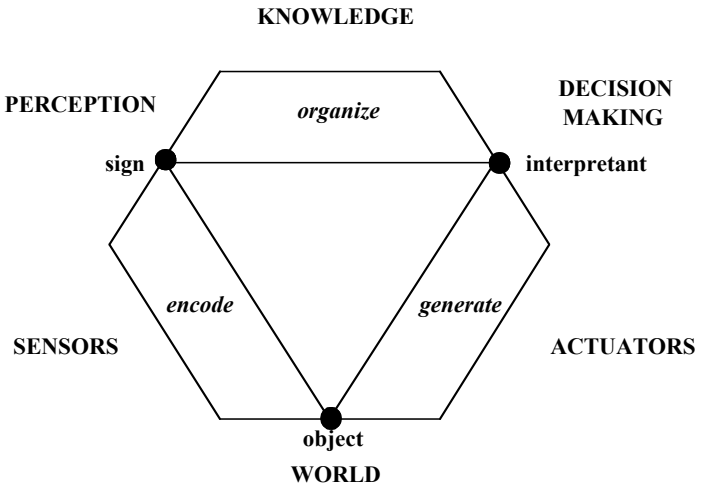


Figure 2.1: Relationship between the three domains of semiotics (i.e., sign, object, and interpretant) and the six-box diagram for modeling intelligent behavior and thinking (e.g., world, sensors, perception, knowledge, decision making, and actuators).

The basic principle of multi-resolution in 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 (e.g., we focus attention on a region in an image). Then,

different combinations of this data are generated based on some similarity criteria (e.g., pixels in the region of interest are grouped based on color, texture, etc). The combination or grouping providing the best results generates data at the next level (e.g., pixel groupings that correspond to sub-regions of objects). 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 (see discussion on [93]). 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 [25][74][139]. Del Bimbo [25] applies the semiotics idea of producing meaning at two levels, the narrative level and discourse level, to automatically classify, annotate, and retrieve videos of commercials using compositional semantics. The narrative level includes basic signs and the values of sign combinations (e.g., color, motion, shapes, feature changes through time); the discourse level describes how to use narrative elements to create a story (e.g., characters, roles, and actions, among others, that provoke some visual message and emotions).

Smoliar et al. [139] describe some of the implications to multimedia search from the point of view of writing and reading multimedia signs. Multimedia such as images and words are considered to signify notions of objects in the world; for example, both an image of a carrot and the word "carrot" signify the notion of carrot. Search is identified as fundamental for the processes of reading and writing so the authors use semiological

ideas in searching for signs such as specifying the information plane of the search as data derived from images or objects depicted in images.

Joyce et al. [74] propose a semiotics framework for integrating high-level metadata (e.g. "carrot") and low-level features (e.g. color histogram extracted from an image of a carrot) by formally adding a second representation level to Smoliar et al. [139]. In this level, features extracted from multimedia are considered signs of multimedia (see Figure 2.2). Textual signs 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 [83][144][145] and neural-network classification agents [74].

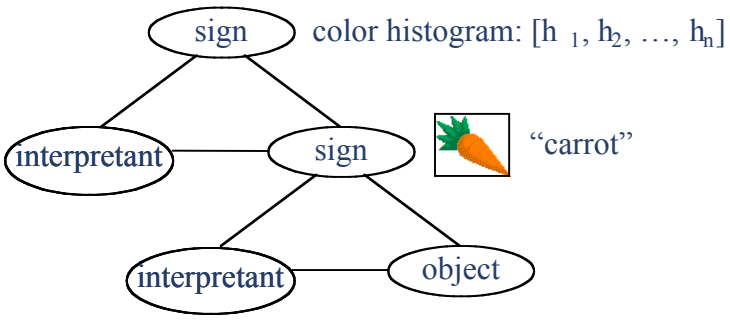


Figure 2.2: Semiotic framework for multimedia and features extracted from multimedia proposed by Joyce et al. [74].

2.3 Semantic Frameworks

Semantic frameworks use text (i.e., words) to represent knowledge about the world. Related work is traditional textual thesauri [60][146][115], WordNet [95], Cyc [39], and the VISAR system [35], among others.

2.3.1 Traditional Textual Thesauri

Traditional textual thesauri consist of terms and relationships between terms. A term is a word or a group of words representing objects, situations, and events in the world. Typical relationships among terms in thesauri are the relationships: "equivalent", "narrower", "broader", and "related". Two terms are equivalent if they represent the same or very similar concepts (e.g., "human" and "man"). A term is narrower than another term if the former concept is a specialization of the latter concept (e.g., "animal" and "dog"). "broader" is the inverse relationship of "narrower". Finally, two terms are related if they have a relationship other than equivalence, generalization, and specialization.

Textual thesauri for a domain are usually constructed manually by experts in the domain such as the INSPEC Classification and Thesaurus [60], the Art and Architecture Thesaurus [146], and the Thesaurus for Graphic Materials I/II [115]. Classification schemas are similar to thesaurus and are also built manually. For example, the Dewey Decimal Classification (DDC) schema was conceived by Melvil Dewey in 1873 and has become the most widely used library classification system in the world.

Thesauri can be built automatically from a collection of textual documents by estimating the similarity between terms based on co-occurrence information. In this case, the thesaurus only has association relationships. As an example, the association thesaurus [69] is constructed automatically based on the co-occurrence of pairs of terms and

phrases; a phrase is a sequence of terms that can represent a concept. Another example is WEBSOM [59][78], which finds associations among words based on immediately surrounding or nearby words in the textual documents using self-organizing maps [77].

Textual thesauri are commonly used in information retrieval systems to modify and expand keyword queries for improved retrieval effectiveness [69]. Some systems automatically build thesaurus-like structures for browsing textual documents [50][51][59][78][169].

2.3.2 WordNet

WordNet [95] is an electronic dictionary that organizes English words into sets of synonyms (e.g., "rock, stone") and connects them with semantic relationships (e.g., "specializes", "contains", and "entails").

Each synonym set in WordNet represents an underlying lexical concept. The semantic relationships incorporated by WordNet are synonymy (similar), antonymy (opposite), hypernymy/hyponymy (specializes/generalizes), meronymy/holonymy (component/ composed of), troponymy (manner of), and entailment (cause, consequence, or involved by necessity).

WordNet is like a manually-constructed general-purpose semantic network and thesaurus with additional types of relationships. WordNet has been extensively used in information retrieval [53][55][86][117], word-sense disambiguation [4][119][141][167], and, even, image retrieval and browsing [2][165], among others.

2.3.3 Cyc and VISAR System

Cyc [39] aims at representing and reasoning on everyday life knowledge. Several image retrieval systems use different components of Cyc to search for images based on textual captions and queries.

Cyc represents facts about everyday life using the Cyc representation language [40], which is based on First-Order Logic. Experts have entered facts manually in Cyc for years; however, Cyc also has an inference engine to deduce new facts from previous ones. Facts about the world are grouped and positioned in conceptual contexts [82], which are independent conceptual dimensions (e.g., absolute time – Jan 1, 2000 -, type of place – outdoors -, and topic – business -), to enable efficient entry and access of knowledge. Cyc has been used for retrieving photos and other captioned data using textual queries by processing the textual captions associated with the photos with the Cyc natural-language processor and expressing them using the Cyc representation language [41]. Textual queries are processed in the same way. No implementation details or evaluation experiments are provided for this application.

In a similar way to [41], the VISAR system [35] expresses the textual captions of documents using Cyc's representation language. In addition, it merges textual queries with the concept network representation of the user's focus of attention in previous queries, and retrieves the citations that match the resulting concepts in order of priority. In the VISAR system, users can also browse and modify the concept network

representations of queries and retrieved documents. These two representations are networks of nodes and arcs representing the most relevant concepts and their associations in the query and the retrieved documents, respectively. The concept network representations displayed to the user have a limited number of nodes, seven nodes, based on the size of the human short-term memory [96].

2.4 Perceptual Frameworks

Perceptual frameworks use audio and/or visual data (e.g., images and texture features) to represent knowledge about the world. Related work is the Texture Thesaurus [85] and the Visual Thesaurus [114], among others.

2.4.1 Texture Thesaurus

The Texture Thesaurus [85] is a two-level thesaurus of texture patterns created using Gabor texture features. The texture patterns in the first level of the thesaurus are obtained by a learning similarity algorithm based on self-organizing maps [77] and linear vector quantization [77]. The second level is constructed using a hierarchical vector quantization technique. The relationship between texture patterns in the first and second levels is similar to relationship "specializes"/"generalizes".

The texture thesaurus has been proven to provide efficient indexing and retrieval of large aerial photographs while maintaining acceptable retrieval effectiveness. After segmenting the images into homogeneous regions, the regions are indexed by the best-

matched texture pattern in the thesaurus. During query time, the regions indexed by the most similar texture pattern to a query are ranked using the Euclidean distance to the query; the best matches are returned as the results of the query.

2.4.2 Visual Thesaurus

The Visual Thesaurus [114] is composed of visual data and relationships among the visual data. Visual data are portions of images and video programs. The Visual Thesaurus considers and adapts the typical thesaurus relationships to the visual domain. For example, two images are visually equivalent if one is the gradual zoom or pan of the other (i.e., equivalent images look essentially the same); one image is visually narrower than another if the former depicts the same item over a much closer scale than the latter (e.g., pictures of a tree taken from very different distances); finally, two images are visually related if their relationship is of a different nature, for example, similar brightness level. This work also describes a way to extend the visual thesaurus framework to other modalities such as audio.

The proposed approach for constructing the Visual Thesaurus [97] starts with creating region groupings within and across images using different visual features. The system then iteratively learns through user feedback how to combine and select groupings using the AQ learning algorithm and self-organizing maps [77], respectively.

2.5 Integrated Frameworks

Integrated frameworks use text together with audio and/or visual data to represent knowledge about the world. Related work is the Multimedia Thesaurus [144][145], the Mirror System [43], and joint text-visual clustering approaches (Semantic Clustering [4] and probabilistic clustering [4][5]), among others.

2.5.1 Multimedia Thesaurus

The Multimedia Thesaurus (MMT) [144][145] is a network of concepts, semantic relationships between concepts, and media representation of concepts. The concept network in Figure 1.2 could be part of a Multimedia Thesaurus. Concepts in the MMT are abstractions of semantically meaningful objects in the real world so texture patterns such as the ones in the Texture Thesaurus [85] are excluded. Concepts are represented by portions of multimedia and by features extracted from the multimedia. However, telationships between concepts in the Multimedia Thesaurus are restricted to the typical thesaurus relationships (i.e., equivalence, narrower/broader, and related). Therefore, perceptual relationships and media examples of relationships are not supported. An example of a perceptual relationship between two concepts is the similar shape relationship; a media example or representation of this relationship could be a condition of a shape feature similarity (see Figure 3.2).

A Multimedia Thesaurus is constructed for a collection of annotated art images in a semi-automatic process [144][145]. The concept network is a manually selected subset of

the Dewey Decimal Classification (DDC) schema. Annotated images are connected to concepts by matching the textual annotations of the images with the textual descriptions of the concepts using latent semantic indexing on a large corpus of art documents. Images without annotations are connected to concepts linked to visually similar images. Another proposed approach to detect concepts in new images is a system of multiple autonomous neural-network classification agents specialized in certain concepts and features [83]. However, none of the two approaches for detecting concepts in images exploits the concepts or the topology of the Multimedia Thesaurus. In other words, the fact that concept "pedigree" is a specialization of concept "dog" is not taken into account when detecting the concepts in an image. Therefore, the examples of concept "pedigree" are not used as examples of concept "dog" in training the corresponding classifier.

The MAVIS 2 system [144] uses the Multimedia Thesaurus to extend queries with equivalent representations in different media and to limit the scope of a query to selected concepts in the Multimedia Thesaurus. The MAVIS 2 system also provides limited capabilities for browsing the Multimedia Thesaurus. For each concept, the interface displays the directly broader and narrower concepts, the media representations, and the documents linked to the concept. The concepts are displayed on the screen in a pre-defined arrangement without the use of advanced visualization techniques.

2.5.2 Mirror System

The Mirror System [43] includes a thesaurus that consists of textual terms, image region clusters, and associations between them. The relationships between terms and region clusters are generic associations with assigned weights. Therefore, no specific semantic or perceptual relationships can be represented between concepts (i.e., terms and region clusters) in the Mirror System.

The Mirror System uses annotated images to build the thesaurus. The words in the annotations become the terms of the thesaurus; no word-sense disambiguation is performed. Images are segmented into regions that are clustered based on different visual features. Associations between terms and region clusters are obtained based on co-occurrence statistics of words and region clusters in images. The Mirror System does not reuse or incorporate any knowledge from external resources such as WordNet. Knowledge extraction is an expensive process that calls for the reuse of available resources.

The Mirror System allows users to retrieve images using textual queries. The Bayesian retrieval model in INQUIRE [152] is applied to (1) translating textual queries into visual queries using the associations between terms and region clusters, and (2) matching visual queries to the images in the database. The Mirror system expects to incorporate relevant feedback from users to refine the associations between terms and regions clusters and to improve the matching of visual queries in the future.

2.5.3 Semantic Clustering

The Semantic Clustering approach [127] defines a hierarchy of semantic classes whose leaf nodes can be represented by different visual features called templates. Semantic classes and feature templates can be considered equivalent to semantic and perceptual concepts, respectively. However, this approach does not support non-strictly hierarchical or perceptual relationships between concepts. Actually, the meaning of the hierarchical relationship is not defined; it is not clear if it is a generalization, composition, or another kind of relationship.

Examples of semantic classes in the Semantic Clustering approach are application classes such as "Satellite" and "Geographical", and semantic clusters such as "Water" and "Residence". The semantic classes are manually defined and organized in a hierarchy by experts. Feature templates are centroids resulting from clustering training images in semantic clusters. Other images are assigned to the semantic clusters if they are visually similar to the feature templates. The Semantic Clustering approach does not reuse or incorporate knowledge from external resources either.

The hierarchy of semantic classes and feature templates in the Semantic Clustering approach is used for efficient indexing and retrieval of images using different visual features. Visual features extracted from visual queries are matched to the feature templates of the semantic clusters; different strategies are proposed to retrieve images by

applying AND or OR logical functions to features during queries. No visualization of images or query results is considered in this work.

2.5.4 Probabilistic Clustering

The probabilistic clustering approach [4][5] also clusters images hierarchically, in this case, by learning the joint distributions of words and visual features using Hofmann's hierarchical clustering model [58]. The resulting structures are hierarchies of clusters (i.e., concepts) that are represented by a combination of words and visual features. The limitations of this work are similar to the Semantic Clustering [127]: it cannot represent non-strictly hierarchical or perceptual relationships between concepts; and the meaning of the hierarchical relationships is not defined. The resulting probabilistic clusters are used to automatically annotate new images and regions in images with words.

2.6 Summary

This chapter has reviewed relevant literature on knowledge representation, discovery, and application using multimedia. Knowledge is usually defined as facts about the world and represented as nodes and arcs between the nodes. Nodes represent objects, concepts, events, or situations (e.g., watermelon and fruit), and the links represent relationships between nodes (e.g., watermelon is a fruit).

Traditional knowledge representation models and frameworks use only text to represent semantic information about the world. However, there are frameworks that propose to

represent perceptual information using images and/or audio. More recent frameworks use any media and features extracted from the media to represent either perceptual or semantic knowledge about the world. However, existing frameworks still lack the expressive power to represent the rich knowledge in human mental models, which contain interconnected nodes of textual and audio-visual nature, as we discuss in chapter 3.

3 Multimedia Knowledge Representation

3.1 Introduction

In this chapter, we present a novel framework for representing knowledge using multimedia, MediaNet. Existing frameworks lack the expressive power to represent the rich knowledge in human mental models.

As mentioned before, knowledge can be represented semantic networks of nodes and arcs [116]. Nodes represent objects, concepts, events, or situations (e.g., watermelon and fruit), and the links represent relationships between nodes (e.g., watermelon is a fruit). Knowledge can refer to either perceptions (the perception of things, e.g., green pattern) or semantics (the meaning of things, e.g., meaning of word "watermelon") in the world so we distinguish two different kinds of knowledge, perceptual and semantic knowledge, respectively. We believe that perceptual and semantic knowledge need to be unified in the same representation framework to impact multimedia information systems, which have to deal with multimedia at the semantic and perceptual levels, as humans do.

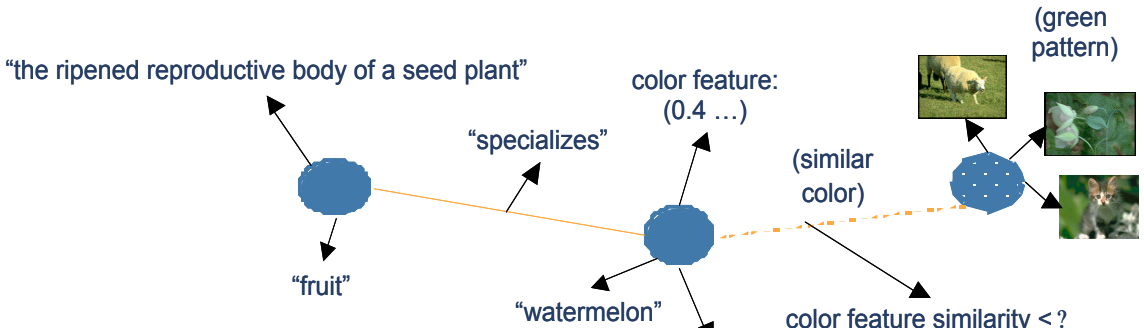


Figure 3.1: Example of a medianet that represents concepts "fruit", "watermelon", and green pattern (circles) illustrated by text, an image region, images, and a color feature. Semantic relationship "specializes" relates concepts "fruit" and "watermelon". Concepts "watermelon" and green pattern are related by similar color perceptual relationship, which is exemplified by a condition on a color feature similarity.

MediaNet is an integrated knowledge representation framework, the first of its kind to use multimedia for representing generic perceptual and semantic knowledge about the world [22][23]. The main components of the MediaNet framework include concepts, relationships between concepts, and media exemplifying concepts and relationships such as images, words, and low-level features of the media. In other words, the MediaNet framework represents knowledge as concept networks with media examples, which we refer to as "multimedia concepts networks" or "medianets". Figure 3.1 shows an example of a medianet that illustrates concepts "fruit", "watermelon", and green pattern (circles) illustrated by text, an image region, images, and a color feature. Semantic relationship "specializes" relates concepts "fruit" and "watermelon". Concepts "watermelon" and "green pattern" are related by similar color perceptual relationship, which is exemplified by a condition on a color feature similarity. The MediaNet framework extends and differs from related work such as the Multimedia Thesaurus [144] and the Visual Thesaurus

[114] in two ways: (1) in combining concepts and relationships at the perceptual and semantic levels in the same network, and (2) in illustrating perceptual and semantic relationships using media.

In designing the MediaNet framework, we have built on the basic principles of semiotics [93] and semantic networks [116], in addition to utilizing evidence from psychology that humans have rich mental models of the world that contain interconnected nodes of textual and audio-visual nature [44][70][79][123]. The MediaNet framework offers functionality similar to that of a dictionary, ontology, or thesaurus by defining and describing semantic concepts. The MediaNet framework extends these frameworks by denoting concepts and relationships between concepts at the semantic and perceptual levels using multimedia. We use MPEG-7 tools [61][87] to encode, represent, and exchange medianets in a machine processable, reusable, and interoperable form. Medianets are encoded using MPEG-7 tools that describe structured collections [133] (e.g., collections of images), the structure [19] (e.g., regions in images), and the semantics [19] (e.g., people depicted in an image) of multimedia (see appendix 11.1).

3.1.1 Outline of the Chapter

The rest of the chapter is organized as follows. In the next section, we review the most relevant work on using multimedia to represent knowledge. In section 3.3, we present evidence from psychology on multimedia mental models in humans. In section 3.4, we describe the main components of the MediaNet framework. In section 3.5, we discuss

the mapping of the MediaNet framework to semantic networks, semiotics, and MPEG-7. Finally, we conclude with a summary of the chapter and a discussion of future work in section 3.6.

3.2 Related Work

This section highlights the limitations of the prior integrated frameworks, extensively reviewed in section 2.5, in comparison with the MediaNet framework. Chapter 2 also reviews other relevant literature including strictly semantic and perceptual frameworks.

As the MediaNet framework, the Multimedia Thesaurus (MMT) [144][145] is a network of concepts and relationships between concepts. However, concepts and relationships are limited to the semantic level. In the MMT, concepts are abstractions of semantically meaningful objects in the real world (e.g., watermelon) so unnamed perceptual patterns (e.g., green pattern) are excluded. Relationships between concepts in the MMT are also restricted to the typical semantic relationships in thesauri (i.e., equivalence, narrower/broader, and related). In addition, only concepts can have media examples in the MMT. In the MediaNet framework, relationships between concepts can be exemplified using multimedia too (e.g., see similar color relationship in Figure 3.1).

The Mirror System [43] includes a thesaurus that consists of textual terms, image region clusters, and associations between them. Terms and region clusters can be considered equivalent to semantic and perceptual concepts, respectively. This thesaurus has several limitations. First, examples of semantic concepts are restricted to words. Second, this

thesaurus does not support any relationships between semantic concepts, or between perceptual concepts. Finally, associations between terms and region clusters cannot have media examples; these relationships can only be characterized with weights.

The Semantic Clustering [127] defines a hierarchy of semantic classes whose leaf nodes can be represented by different visual features called templates. Let's consider the semantic classes and the feature templates equivalent to semantic and perceptual concepts, respectively; then, this approach does not support non-strictly hierarchical or perceptual relationships between concepts. Actually, the meaning of the hierarchical relationships between concepts is not defined. It is not clear whether they are generalization or composition relationships. As for the Mirror System [43], media examples of semantic concepts are restricted to words.

As the Semantic Clustering, Barnard et al. [4][5] also cluster images hierarchically, in this case, by modeling the joint distribution of words and visual features. The resulting structures are hierarchies of clusters (i.e., concepts) that are represented by combinations of words and visual features. The limitations of this work are similar to the Semantic Clustering [127]'s: it cannot represent non-strictly hierarchical or perceptual relationships between concepts; and the meaning of the hierarchical relationships is not defined, among others.

The MediaNet framework extends and differs from related work in two ways: (1) in combining concepts and relationships at the perceptual and semantic levels in the same network, and (2) in illustrating perceptual and semantic relationships using media.

3.3 Multimedia in Human Mental Models

The goal of psychology is to understand human intelligence. Two important trends can be distinguished in psychology: behaviorism and cognitive psychology. Behaviorism [129] 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 [26] 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 knowledge representation and semiotics, among others.

There is ample support from psychologists, especially cognitive psychologists, for the notion that human mental models of the world contain nodes of not only textual (semantic) but also audio-visual (perceptual) nature [44][70][79][123]. For example, Rumelhart and Norman [123] argue that semantic memory contains nodes that are not named after words in some language. Both Kosslyn [79] and Johnson-Laird [70] agree that mental models represent information using textual propositions and images. In addition, Johnson-Laird [70] acknowledges the nature of some mental representations is temporal and dynamic, e.g., images in motion. Regarding audio, several music cognition studies such as Downling and Harwoord [44] have provided strong evidence for long-term auditory memory.

The MediaNet framework builds on this work by using multimedia to represent generic perceptual and semantic knowledge in advanced multimedia information systems. We draw the parallel between multimedia and human sensory data [134].

3.4 The MediaNet Framework

This section describes the main components of the MediaNet framework: (1) concepts, (2) relationships between concepts, and (3) media examples.

The MediaNet framework represents knowledge using concepts, relationships, and media examples. In the MediaNet framework, concepts can represent either semantically meaningful entities (e.g., objects, events or abstract concepts) or perceptual patterns (e.g., color pattern) in the world. Concepts are defined and/or exemplified by multimedia such as images, video, audio, text, and audio-visual features. The MediaNet framework models traditional semantic relationships between concepts such as "specializes" and "contains" but also adds additional functionality by modeling perceptual relationships based on feature similarity and constraints (e.g., condition on the distance of color histograms).

An example of a medianet with several concepts and relationships at the semantic and perceptual levels, and diverse media examples is shown in Figure 3.1. Figure 3.2 depicts another example medianet that represents concepts "human" and "hominid" illustrated by text, an image region, audio, and a shape feature. The two concepts are related by semantic relationship "specializes" and similar shape perceptual relationship; the latter relationship is exemplified by a condition on a shape feature similarity.

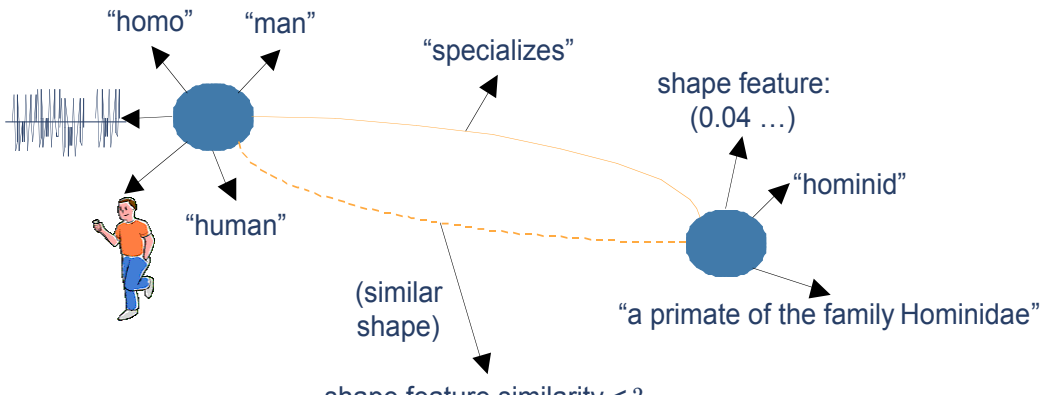


Figure 3.2: Example of a medianet that represents concepts "human" and "hominid" (circles) illustrated by text, an image region, audio, and a shape feature. The two concepts are related by semantic relationship "specializes" and similar shape perceptual relationship; the latter relationship is exemplified by a condition on a shape feature similarity.

3.4.1 Concepts

Concepts are the basic units for knowledge representation in the MediaNet framework. Concepts refer to entities in the world under representation. More specifically, concepts are the notion or meaning, interpretant in semiotics terms (see section 2.2.2), of world entities. We distinguish between two different kinds of concepts, semantic and perceptual concepts, based on whether the concepts have textual representations or not, respectively.

Semantic concepts refer to entities named with words in a language (see concept "watermelon" and concept "human" in Figure 3.1 and Figure 3.2, respectively). Examples of semantic concepts are the semantic entities identified by MPEG-7 for describing the semantics of multimedia, i.e., objects, events, places, times, states, and

abstract concepts [19]. Objects and events are entities that exist or take place in time and space in the world. Objects can be, among others, living entities such as people, animals, and plants; man-made objects such as vehicles, buildings, and furniture; and natural objects such as mountains, rivers, and stars. Examples of events are actions (e.g., run and fly), social events (e.g., party and sport game), and natural phenomena (e.g., sunset and rain). Countries and years are examples of places and times, respectively. States are properties of entities such as the weight of a table or the cloudiness of a day. Abstract concepts are entities with no physical presence such as happiness and freedom. Semantic concepts can refer to classes of semantic entities in the world such as "car"; or to specific and unique entities such as "Ronal Reagan".

Perceptual concepts refer, instead, to entities defined using perceptual features and representations (see green pattern concept in Figure 3.1). For humans, perceptual concepts can be visual and audio patterns, for which we do not have words. For example, a visual pattern can be the color of specific fabrics; an example of an audio pattern is the rhythm of some songs by the same artist. In terms of multimedia, perceptual concepts can be clusters of multimedia based on low-level features extracted from the media. A group of images with similar color histograms is an example of a perceptual concept.

It is important to point out the differences between words in a language and concepts, especially semantic concepts. A concept may be represented by more than one word and a word may designate several concepts. The former case corresponds to synonyms in a language, i.e., words having the same or nearly the same meaning in a language such as

"human" and "man". The latter case amounts to polysemy, i.e., a word having multiple meanings in a language, such as "plant" which can be a "building for carrying on industrial labor" or "a living organism lacking the power of locomotion". As mentioned above, human languages may not have words to designate some concepts, in particular, perceptual concepts.

3.4.2 Relationships Between Concepts

Relationships between concepts represent interactions between concepts based on meaning or perceptual attributes. We, therefore, distinguish between semantic and perceptual relationships. Relationships are the notion or meaning, interpretant in semiotics terms (see section 2.2.2), of interactions between world entities.

Semantic relationships relate concepts based on their meaning (see relationship "specializes" in Figure 3.1 and Figure 3.2). The relationships in traditional thesauri and the electronic dictionary WordNet [95] are examples of semantic relationships; these are listed with definitions and examples in Table 3.1. The typical relationships among terms in a thesaurus are the equivalent, narrower/broader, and related relationships. The relationships in WordNet are synonymy, atonymy, hypernymy/hyponymy, meronymy/holonymy, troponymy, and entailment. A concept usually has only one broader concept so relationship "specializes"/"generalizes" is used to organize concepts hierarchically in traditional thesauri including WordNet.

Table 3.1: Examples of semantic relationships with definitions and examples in traditional thesauri and WordNet.

Traditional Thesauri		
Relationship	**Definition**	**Example**
Equivalent	Equivalent	human ↔ man
Broader Narrower	Generalizes Specializes	fruit → watermelon man → hominid
Related	Associated	walk → run
WordNet		
Relationship	**Definition**	**Example**
Synonymy	Equivalent	rock ↔ stone
Antonymy	Opposite	white ↔ black
Hypernymy Hyponymy	Generalizes Specializes	animal → dog rose → flower
Meronymy Holonymy	Is contained in Contains	ship → fleet martini → gin
Troponymy	Manner of	whisper → speak
Entailment	Causes or requires	divorce → marry

Concepts may refer to entities that can be perceived by the senses. Therefore, concepts can also be related based on the perceptual attributes of the entities they represent (see similar color and shape relationships in Figure 3.1 and Figure 3.2, respectively). Examples of perceptual relationships are visual relationships (e.g., Watermelon has similar shape to Cantaloupe) and audio relationships (e.g., Stork sounds similar to Pigeon). Both semantic

and perceptual relationships can be illustrated using multimedia as detailed in the next section.

The relationships in the MediaNet framework are binary relationships meaning they have one source node and one target node. More general relationships such as N-ary relationships can be represented in terms of several binary relationships. The MediaNet framework can characterize properties of relationships to support inference and reasoning. The most important properties of a binary relationship are reflexivity, symmetry, and transitivity. Let's assume that aRb means that node a is related to node b by relationship R. Relationship R is reflexive if for any node a, aRa. A relationship R is symmetric if aRb implies that bRa. Finally, a relationship R is transitive if aRb and bRc imply that aRc.

3.4.3 Media Examples

Concepts and relationships, which refer to entities and interactions in the world, can be defined and/or exemplified by multimedia such as words, images, and low-level features of the media. Media examples are, therefore, signs in semiotics terms (see section 2.2.2) of world entities or interactions. Figure 3.1 and Figure 3.2 show several media examples of concepts and relationships (see the arrow lines).

Media examples of concepts can be whole multimedia such as entire images, or portions of multimedia such as regions in images. As an example, concept "human" in Figure 3.2 is illustrated using only the region in the image that depicts the human. Concepts,

furthermore, can be exemplified using feature values or models extracted from multimedia examples. Going back to the example in Figure 3.2, the value of a contour shape feature of the image region depicting the human can also be a media example of concept "human". The contour shape feature of the concept could have been obtained by averaging or deriving statistics from the contour shape features of several image regions depicting humans. However, the image regions per se do not need to appear as examples of the concept, only the contour shape feature model. Although, images were used to illustrate the fact that portions of multimedia and features extracted from multimedia can be examples of concepts, the same applies to other media such as text, audio, and video.

Relationships between concepts can also have media examples. For example, semantic relations can be exemplified using words such as "specializes" and "contains". Semantic and perceptual relationships related to audio and visual relationships can be represented by other media. For example, a similar shape relationship can be exemplified as a condition on a shape feature similarity. This and other examples of semantic and perceptual relationships with their media examples are shown in Figure 3.1 and Figure 3.2.

It is important to point out that some media examples may be more relevant and applicable than others to specific concepts and relationships. For example, cars can be of many colors; therefore, a color feature is not very representative or relevant of concept "car". Moreover, audio examples do not apply to concept "sky". In the MediaNet

framework, weights specifying relevance and strength can be assigned to the media examples of concepts.

Figure 3.3 shows additional media examples of several concepts. Concept "car" is associated with the word "car", the textual definition "4-wheeled motor vehicle; usually propelled by an internal combustion engine", an image depicting a car together with shape features extracted from the image, and the sound recording of a running car. Concept "blue" can have the English word "blue", the Spanish word "azul", the textual definition "the pure color of a clear sky; the primary color between green and violet in the visible spectrum", and a color histogram corresponding to standard blue color. Textual examples are not restrictive to any language in particular.

Figure 3.3: Media examples of concepts "car" (left) and "blue" (right).

3.5 Mapping to Semantic Networks, Semiotics, and MPEG-7

In designing the MediaNet framework, we have built on the basic principles of semiotics [93] and semantic networks [116]. We use MPEG-7 structured collection, structure, and semantic description tools to encode and represent medianets in a machine processable,

reusable, and interoperable form. This section discusses the mapping of the MediaNet framework to semantic networks, semiotics, and MPEG-7.

3.5.1 The MediaNet Framework as a Semantic Network

Semantic networks [116] use nodes to represent objects, concepts, or situations; and arcs to represent relationships between nodes (e.g. the statement "human specializes homonid" could be represented by the chain: node "human" – arc "specializes" – node "homonid"). Section 2.2.1 provided an overview of semantic networks and other knowledge representation models.

The components of the MediaNet framework can be mapped to semantic networks although with certain extensions (e.g., having relationships as nodes). Concepts and media examples are nodes in semantic networks. Relationships between concepts in the MediaNet framework need to be nodes in semantic networks because they can have media examples. Media examples and corresponding concepts or relationships are linked by arcs of type "media example of". Arcs of type "source of" and "target of" link relationship nodes to the source and target concept nodes, respectively. If a relationship does not have any media examples, it could be represented as a simple arc in the semantic network. The medianet shown in Figure 3.2 is represented as a semantic network in Figure 3.4.

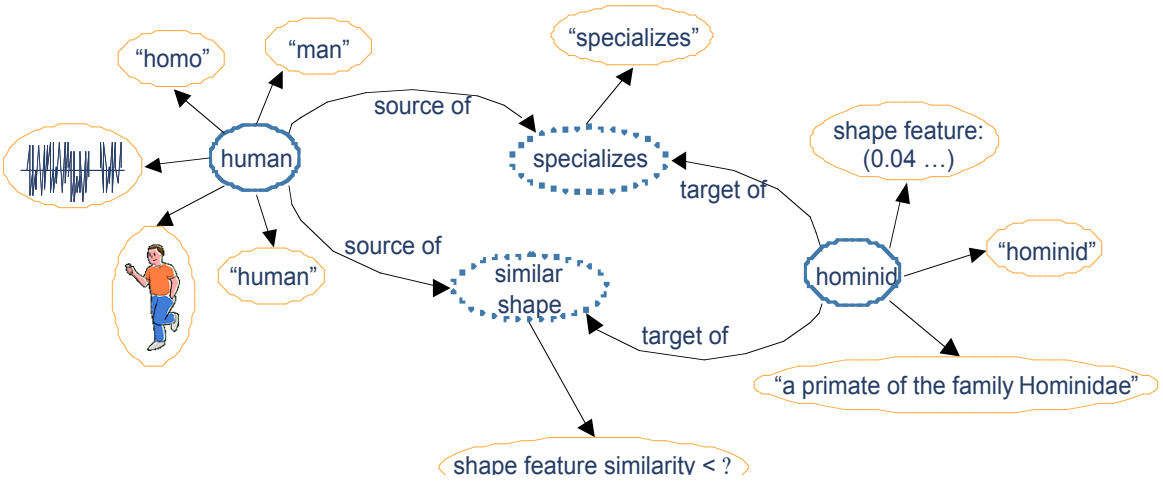


Figure 3.4: Semantic network corresponding to the medianet shown in Figure 3.2. The thick solid and dashed circles are nodes in the semantic network that correspond to concepts and relationships in the medianet, respectively. The thin circles are nodes that represent media examples of concepts and relationships. The arrow lines are arcs in the semantic network. The straight arcs are of type "media example of". The type of the curved arcs is specified on the figure, either "target of" or "source of".

3.5.2 Semiotics View of the MediaNet Framework

As presented in section 2.2.2, semiotics [93] 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 focus on the two basic principles of semiotics that are most useful for multimedia information systems. The first principle is the decomposition of semiotics into three domains: syntax (sign; word "car"), semantic (interpretant; notion of car), and pragmatics (object; real object car). The second principle, the multi-resolution principle, arises from modeling a unit of intelligence as three cognitive processes applied repeatedly: focusing attention, combinatorial search, and grouping (or generalization).

The entities and their interactions in the world represented by concepts and relationships in the MediaNet framework, respectively, are objects at the pragmatic level of semiotics. Both entities and interactions are considered objects because they exist in the world although they might not be tangible (e.g., "happiness" and "specializes"). We also draw the parallel between concepts and relationships in the MediaNet framework, and interpretants at the semantic level of semiotics. The media examples including multimedia and features extracted from multimedia correspond, respectively, to signs of the objects and of multimedia at the syntax level (as proposed in [74]). Figure 3.5 illustrates the application of the semiotic framework of object, interpretant, and sign to concept "human" for the medianet in Figure 3.2 and its corresponding semantic network in Figure 3.4. We assume a shape feature also illustrates concept "human". Finally, relationship "specializes"/"generalizes" in the MediaNet framework and the summarization of medianets (see section 4.5) support the definition of hierarchies of concepts at multiple levels implementing the semiotic principle of multi-resolution.

An interpretant is conferred upon a sign when a user interprets the sign (media). The interpretation of a sign, therefore, may be relative to users and the tasks under execution, as claimed by Joyce et al. [74]. The MediaNet framework aims at recording some possible, rather than all, interpretations of media to support more advanced classification, browsing, and retrieval functionalities in multimedia information systems. Information about how specific users executing particular tasks may interpret a sign can be included in the MediaNet framework in the form of contextual information similar to

the modality/disposition/epistemology context dimension proposed by Lenat [82], which specifies if a fact represents a belief or a desire, and the people who believe in it.

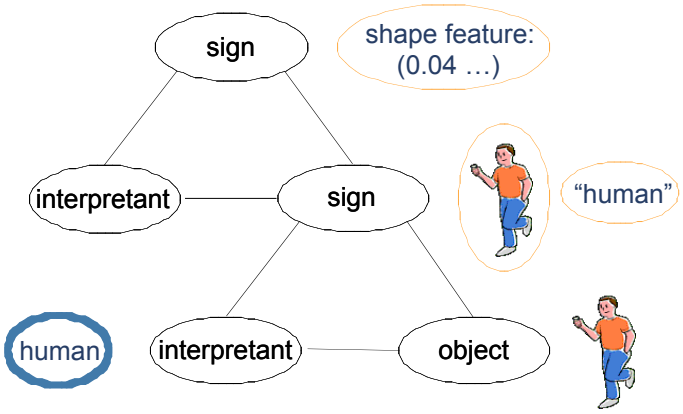


Figure 3.5: Semiotic framework of object, interpretant, and sign applied to concept "human" in Figure 3.2 and Figure 3.4. We assume a shape feature also illustrates concept "human". Each triangle represents the semiotic framework composed by objects, interpretants, and signs. The bottom triangle illustrates a real human (object) represented by the concept human in the medianet (interpretant) and encoded as a region in a image and the word "human" (signs); the top triangle illustrates the image region (object) encoded as a shape feature vector (sign).

3.5.3 MPEG-7 Encoding of the MediaNet Framework

This section presents our approach to encode and represent medianets using MPEG-7 structured collection, structure, and semantic description tools in a machine processable, reusable, and interoperable form.

MPEG-7 has standardized tools for describing different aspects of multimedia [61][87]. It has potential to advance current and new multimedia information applications because it provides a standard, flexible, and interoperable way for describing multimedia [102]. In

particular, the MPEG-7 standard includes tools that describe structured collections [133] (e.g., collections of images), the structure [19] (e.g., regions in images), and the semantics [19] (e.g., people depicted in an image) of multimedia.

The structured collection description tools [133] can describe clusters of multimedia (e.g., collections of images) together with their attributes (e.g., labels and descriptor probability models) and relationships (e.g., cluster A is disjoint with cluster B). The structure description tools [19] describe the structure of multimedia data in space and/or time by describing segments of multimedia (e.g., still regions) together with their attributes (e.g., color features) and the relationships among them (e.g., still region A to the left of still region B). The semantic description tools [19] represent semantic entities (e.g. objects and events) in narrative worlds depicted by multimedia together with their attributes (e.g., labels and definitions) and relationships (e.g., object A is the agent of event B). More information about these description tools is provided in appendix 11.1.

We use structured collection, structure, and semantic description tools to encode and represent medianets, i.e., multimedia concept networks from the MediaNet framework, as described below. The concept network is described using the structure collection description tools; whereas, structure and semantic description tools are used to represent the media examples in the medianet. The MPEG-7 description of the medianet in Figure 3.2 is shown in Table 3.2. For simplicity, no media examples are described for the relationships. We assume readers are familiar with MPEG-7 and the markup language XML [157].

Table 3.2: MPEG-7 description of the medianet in Figure 3.2. For simplicity, no media examples are described for the relationships.

```
<!-- MPEG-7 structured collection description tool -->
<StructuredCollection>
 <!-- Describes concept "human" -->
 <ClusterModel id="CM-Human">
  <!-- MPEG-7 semantic description tool -->
  <Semantics xsi:type="AgentObjectType">
   <Label>  <Name> Man </Name>  </Label>
   <Label>  <Name> Homo </Name>  </Label>
   <Label>  <Name> Human </Name>  </Label>
  </Semantics>
  <Collection xsi:type="ContentCollectionType">
   <Content xsi:type="ImageType">
    <!-- MPEG-7 structure description tool -->
    <Image>
     <MediaLocator xsi:type="MediaLocatorType">
      <MediaUri>Human.jpg</MediaUri>
     </MediaLocator>
    </Image>
   </Content>
   <Content xsi:type="AudioType">
    <!-- MPEG-7 structured collection description tool -->
    <Audio>
     <MediaLocator xsi:type="MediaLocatorType">
      <MediaUri>Human.wav</MediaUri>
     </MediaLocator>
    </Audio>
```

```
    </Content>
  </ClusterModel>
  <!-- Describes concept "hominid" -->
  <ClusterModel id="CM-Hominid">
   <!-- MPEG-7 semantic description tool -->
   <Semantics xsi:type="ObjectType">
    <Label>  <Name> Hominid </Name>  </Label>
    <Label>  <Name>  a primate of the family Hominidae  </Name>  </Label>
   </Semantics>
   <DescriptorModel>
    <Descriptor xsi:type="SomeShapeFeatureType" numOfCoeff="16">
     <Coeff> 6 7 8 9 0 1 2 3 4 5 6 1 2 3 4 5 </Coeff>
    </Descriptor>
    <Field>Coeff</Field>
   </DescriptorModel>
   <ProbabilityModel xsi:type="ProbabilityDistributionType" dim="16">
    <Mean mpeg7:dim="16"> 5 6 7 8 9 0 1 2 3 4 5 0 1 2 3 4 </Mean>
   </ProbabilityModel>
  </ClusterModel>
  <!-- Describes relationships between concepts "human" and "hominid" -->
  <Relationships>
   <Node id="source" href="#CM-Human"/>
   <Node id="target" href="#CM-Hominid"/>
   <Relation type="urn:mpeg:mpeg7:cs:SemanticRelationCS:2001:specializes"
           source="#source" target="#target"/>
   <Relation type="SomeRelationCS:similarShape"
           source="#source" target="#target"/>
  </Relationships>
</StructuredCollection>
```

A medianet is represented using a structured collection in MPEG-7. The concepts in the network are clusters in the structured collection. Textual examples of concepts can be labels of clusters. If textual examples correspond to specific and known semantic entities (e.g., an object or an event), they can be, instead, semantic descriptions of the clusters. On the other hand, non-textual examples, such as images, image regions, and feature values, can be multimedia, segments, and descriptor items of the cluster, respectively. If feature values for concepts were obtained by statistical analysis such as the mean of several feature values, they can be described as descriptor probability models of the clusters. See appendix 11.1 for details.

Relationships among concepts are described as relationships among the corresponding clusters in the structured collection. If a relationship has media examples, the relationship is also represented as a cluster in the structured collection. The cluster and the relationship in the structured collection that represent the same relationship in the MediaNet framework are related using the MPEG-7 relationship "identify". The relationship "identity" specifies that the related items are actually identical. The media examples of the relationship are represented as labels, semantic descriptions, descriptor items, or descriptor probability models of the corresponding cluster (see appendix 11.1), in a similar way as the media examples of a concept.

3.6 Summary

This chapter has presented a novel framework for representing generic perceptual and semantic knowledge using multimedia, MediaNet. The main components of the MediaNet framework include concepts, relationship between concepts, and media exemplifying concepts and relationships such as images, words, and low-level features of the media.

In designing the MediaNet framework, we have built on the basic principles of semiotics and semantic networks, in addition to utilizing evidence from psychology that humans have rich mental models of the world that contain interconnected nodes of textual and audio-visual nature. We have also proposed to use MPEG-7 structured collection, structure, and semantic description tools to encode and represent medianets in a machine processable, reusable, and interoperable form. The concept network is described using the structured collection description tools; whereas, structure and semantic description tools are used to represent the media examples in the medianet.

There is a clear trend of extending the current Web so that information is given well-defined meaning, better enabling the automation of many services, and the cooperation of computers and people. The approach taken by W3C is to promote the use of semantic markup languages such as RDF (Resource Description Framework) [158] and OWL (Web Ontology Language) [160] for publishing and sharing information and

ontologies on the Web. As future work, we plan to define the encoding and representation of medianets using these markup languages.

4 Multimedia Knowledge Discovery

4.1 Introduction

This chapter focuses on the discovery, summarization, and evaluation of multimedia knowledge from annotated images such as image clusters, word senses, and relationships among them. Existing techniques process each media independently or are domain specific so they do not generalize to arbitrary multimedia or knowledge.

The substantial proliferation of multimedia such as annotated images requires tools for discovering useful knowledge from multimedia to enable innovative and intelligent organization, filtering, and retrieval of multimedia. Perceptual knowledge (e.g., image clusters and relationships between them) is essential for multimedia applications because it can be extracted automatically and is not inherently limited to a set of words as textual annotations. On the other hand, semantic knowledge (e.g., word senses and relationships between them) is more powerful for multimedia applications because human communication often happens at this level. However, current approaches for extracting semantic knowledge are, at best, semi-automatic and very time consuming. Furthermore, it is often necessary to summarize multimedia knowledge in order to reduce the size of the knowledge. Hence, ways to quantify the consistency, completeness, and conciseness

of the multimedia knowledge, among others, are essential to evaluate and compare different techniques for knowledge discovery and summarization.

In this chapter, we present novel and automatic methods for discovering perceptual and semantic knowledge from annotated images, and for summarizing and evaluating arbitrary multimedia knowledge [10][14][15][16]. In the multimedia knowledge discovery process, a medianet is built for a collection of annotated images. The extracted medianet consists of concepts (e.g., image clusters and word senses), relationships between concepts (e.g., relationships "visually similar" and "specializes"), and media examples (i.e., images and textual examples of concepts). In contrast to related work such as visual-text image clustering [4][54] and word-sense disambiguation of image annotations [4][119], we propose to integrate features of both images and textual annotations in the multimedia knowledge discovery process. We also propose novel methods for discovering perceptual relationships and semantic concepts from annotated images. The multimedia knowledge summarization process consists of reducing the size of a medianet, in terms of the number of concepts and relationships. The methods for multimedia knowledge evaluation output quality measures of any medianet. The proposed multimedia knowledge summarization and evaluation techniques differ from related work such as EZWordNet [94] and current evaluation of semantic ontologies [52] in being automatic, generic, and applicable to any multimedia knowledge that can be expressed as a set of concepts, relationships between concepts, and media examples, e.g., medianets.

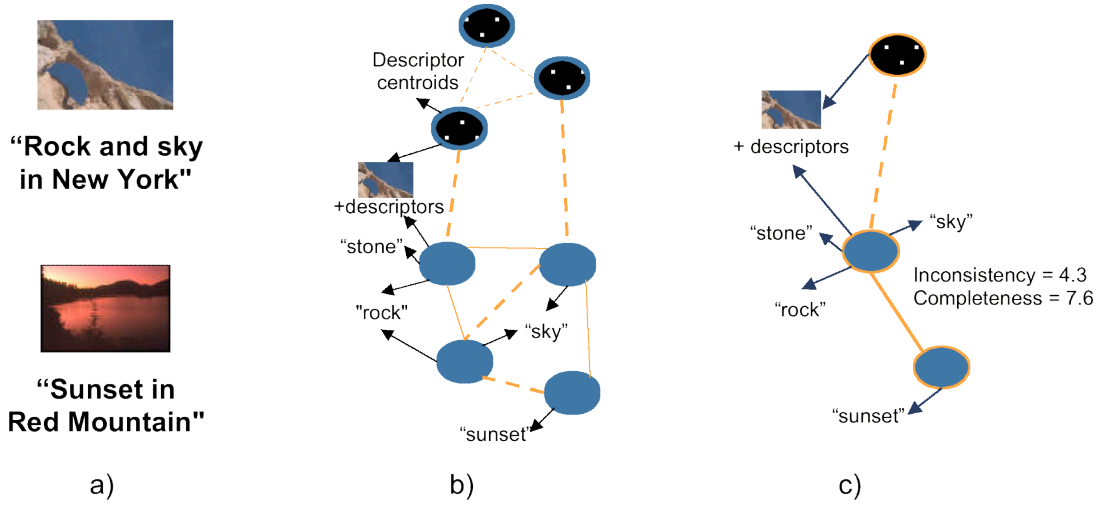


Figure 4.1: Examples of a) typical annotated images, b) a extracted medianet that represents knowledge about the images, and c) a summary of the extracted medianet with some measures of knowledge quality.

The proposed multimedia knowledge discovery, summarization, and evaluation techniques are exemplified in Figure 4.1. Figure 4.1 shows examples of typical annotated images, a medianet that represents knowledge about the images, and a summarized version of the extracted medianet with some measures of knowledge quality (e.g., completeness). We propose to discover perceptual knowledge [10][15] from a collection of annotated images by discovering perceptual concepts by clustering the images based on visual and textual features. We find perceptual relationships among the clusters based on feature similarity and statistical dependencies between the clusters. In the semantic knowledge extraction process [10][16], we discover semantic concepts by disambiguating the sense of words in the annotations using WordNet [95] and the image clusters. We find semantic relationships among the detected senses based on WordNet. The proposed knowledge summarization techniques reduce the size of a medianet (in terms of number

of concepts and relationships) by grouping similar concepts together [10][14]. We calculate the distance or dissimilarity between concepts using a novel concept distance measure based on concept statistics and network topology. In this chapter, we also propose automatic techniques for measuring the consistency, the completeness, and the conciseness of medianets based on information theory and graph notions such as distance spread, entropy, and graph density [10][14].

We have performed several experiments for evaluating the proposed techniques for perceptual and semantic knowledge discovery, and for multimedia knowledge summarization. We evaluate image clusters (perceptual concepts) and word senses (semantic concepts) using a cluster-category correlation measure and the world-sense disambiguation accuracy, respectively. Extracted and summarized medianets were compared with several baseline approaches using the proposed knowledge quality measures (e.g., distance spread, concept entropy, and graph density). Experiments have shown the superiority of integrating the analysis of images and annotations for improving the performance of the image clustering (up to 8% gain in correlation for some categories) and the word-sense disambiguation (6% gain in accuracy for nature images), the importance of good concept distance measures for knowledge summarization, and the potential of automatic measures for evaluating knowledge quality.

4.1.1 Outline of the Chapter

The rest of the chapter is organized as follows. In section 4.2, we review some related work on multimedia knowledge discovery, summarization, and evaluation. In sections 4.3 and 4.4, we present the techniques for discovering perceptual and semantic knowledge from annotated image collections, respectively. In sections 4.5 and 4.6, we describe the proposed methods for multimedia knowledge summarization and evaluation, respectively. In section 4.7, we present the experiments performed for evaluating the proposed techniques. Finally, we conclude with a summary of the chapter and a discussion of future work in section 4.8.

4.2 Related Work

Relevant work on perceptual knowledge discovery includes visual thesaurus construction [85][114], and joint visual-text clustering and retrieval of images [4][54]. The Texture Thesaurus [85] is a perceptual thesaurus limited to texture clusters of regions in satellite images, and is constructed using neural network and vector quantization techniques. The Visual Thesaurus [114] adapts typical concepts and relationships from textual thesauri to the visual domain. An approach for building the Visual Thesaurus involves grouping regions within and across images using visual features. Then, relationships among groupings are learned through user interaction, so it is not a fully automatic system. Barnard et al. [4] clusters images by hierarchically modeling the joint distribution of words and visual features; however, the organization structures are limited to hierarchies.

Grosky et al. [54] uses Latent Semantic Indexing (LSI) [42] and word weighting schemes to retrieve images using concatenated vectors of visual features and category label bits (i.e., bit per category: value 1 when category is present in image and value 0 otherwise). Grosky et al. [54] does not try to discover relevant concepts or relationships from the images and their category labels. However, limited experiments (50 images, 15 categories) have shown some performance improvement in image retrieval.

Relevant work on semantic knowledge discovery includes word-sense disambiguation techniques for textual documents [141][167]. Words in English may have more than one sense or meaning, for example "plant, industrial plant" and "plant, living organism" for the word "plant". Word-sense disambiguation (WSD) is the process of finding the correct sense of a word within a document, which is a long-standing problem in Natural Language Processing. The reason for this is that although most English words have only one sense (80%), most words used in documents have more than one sense (80%) [141]. Actually, it is difficult to do significantly better than choosing the most common sense of each word in spite of much work in the literature [149]. The two principles governing most word-sense disambiguation techniques are (1) that nearby words are semantically close or related and (2) that the sense of a word is often the same within a document [167]. In the literature, there are unsupervised [167] and supervised [141] approaches that often use WordNet as the electronic word-sense dictionary. There are also image indexing approaches that disambiguate the senses of words in image annotations [4][119]. However, none of these approaches combine textual and image features during

the word-sense disambiguation. More recent work than ours proposes to disambiguate words in image annotations using image features [6]. This is a supervised approach that selects the sense with the highest posterior given the word and the image features; therefore, large training sets may be required for extended vocabularies. More recently, there has also been some work on semi-automatic, data-driven construction of multimedia ontologies [65].

Relevant work on multimedia knowledge summarization has been limited to efforts in concept network reduction such as EZWordNet [94] and VISAR [35]. EZ.WordNet.1-2 [94] are coarser versions of WordNet generated by collapsing similar word senses and by dropping rare word senses. This process is governed by five rules manually designed by researchers for WordNet (e.g., collapse two senses of the same word if they have at least two synonyms and they have the same synonyms) so they are not applicable to other kinds of knowledge such as perceptual knowledge or parts of WordNet, which is our case. VISAR [35] is a hypertext system for the retrieval of textual captions. One of the functionalities of the VISAR system is the representation of the retrieved citations as networks of key concepts and relationships. Several reduction operators are used in this process (e.g., replace two concepts with a common ancestor) but the reduction operators are again manually defined and lacking generality.

Finally, prior work relevant on multimedia knowledge evaluation includes manual evaluation of semantic ontologies [52] and automatic but application-oriented evaluation of hierarchical image clusters [5]. Typical criteria used by experts in the evaluation of

semantic ontologies are consistency, completeness, conciseness, sensitiveness, and expandability [52]. Barnard et al. [5] evaluate hierarchical image clusters using an automatic image and region annotation application. The performance of the image annotation is measured by comparing the words predicted by various models using the cluster hierarchy with words actually present in the data. The performance of the region annotation is measured both automatically based on the annotation performance and by manual inspection.

4.3 Perceptual Knowledge Discovery

Our proposed approach for extracting perceptual knowledge from a collection of annotated images consists of three steps, as shown in Figure 4.2 [10][15]. First, visual features and textual features are extracted from the images and textual annotations. Then, perceptual concepts are formed by clustering the images based on their visual and textual features. Finally, perceptual relationships are discovered based on feature similarity and statistical dependencies between the clusters. This section discusses each step in detail.

The methods for image and text processing, and for discovering perceptual concepts, i.e., clustering images, described below are not new. However, we extensively test and report on the effectiveness of different clustering methods and feature combinations for perceptual concept extraction in section 4.7.2. In addition, we focus on the discovery of relationships between the extracted perceptual concepts. If perceptual concepts are

interconnected, then, they can be summarized using the techniques described in section 4.5 by grouping similar clusters.

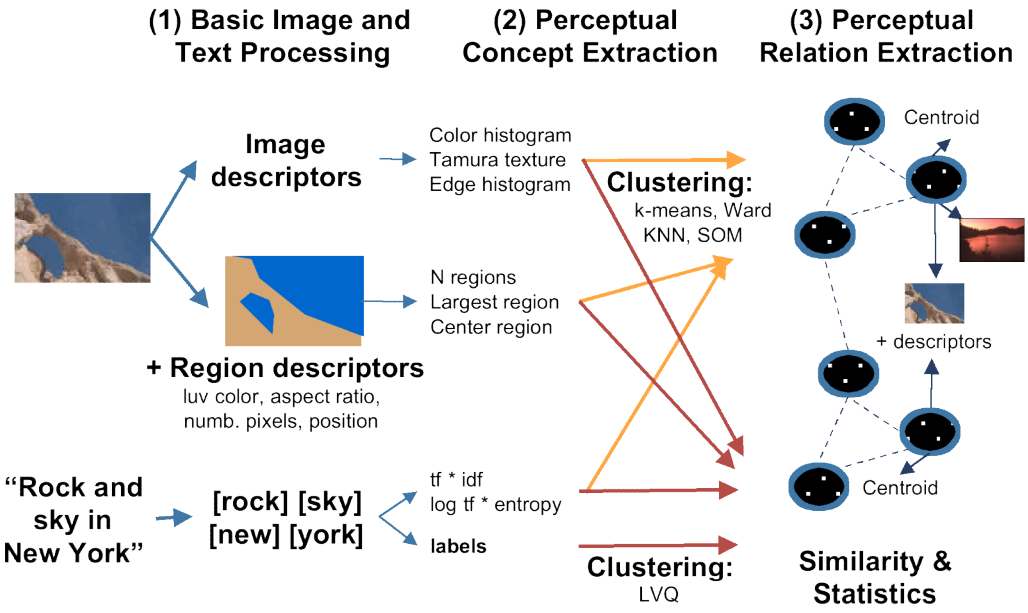


Figure 4.2: Perceptual knowledge extraction process: (1) first, visual and textual features are extracted from the images and the textual annotations, respectively; (2) then, perceptual concepts (dotted ellipses) are obtained by clustering the images based on the features; and, (3) finally, perceptual relationships (dash lines) among clusters are discovered based on cluster similarity and conditional probabilities.

4.3.1 Basic Image and Text Processing

During this step, visual and textual features are extracted from the images and the textual annotations, respectively. Each media is processed independently.

The images are first segmented into regions with homogeneous visual features. Image segmentation consists of grouping visually similar pixels in an image into regions. There are many proposed methods for segmenting images [121]. We use Columbia University's

automatic region segmentation method, which fuses color and edge pixel information [171]. This method has been proven to provide excellent segmentation results. After segmenting the images, features are extracted from the images and the regions for representing visual features such as color, texture, and shape. We use color histogram, Tamura texture, and edge direction histogram globally for images [80][121]; and mean LUV color, aspect ratio, number of pixels, and position locally for segmented regions [171]. This feature set covers the important visual features; moreover, each feature has been independently shown to be effective in retrieving visually similar images or videos in visual databases [80][121][171].

In the basic text processing, the words in the annotations are tagged with their Part-Of-Speech information (POS; e.g., noun and verb). WordNet is used to stem words down to their base form (e.g., "burned" is reduced to "burn") and to correct some POS tagging errors (e.g., "dear" in Figure 4.3 can not be a verb based on WordNet). Then, stop words, (i.e., frequent words with little information such as "be"), non-content words (i.e., words that are not nouns, verbs, adjectives or adverbs such as "besides"), and infrequent words are discarded because of their low relevance. The remaining words for each image are represented as a feature vector using word-weighting schemes [46], which assign weights to words reflecting their discriminating power in a collection. We support (1) tf*idf, term frequency weighted by inverse document frequency; and (2) log tf*entropy, logarithmic term frequency weighted by Shannon entropy of the terms over the documents. The latter has been proven to outperform the former for retrieving textual documents [46].

4.3.2 Perceptual Concept Extraction

The second step is to find perceptual concepts by clustering the images based on visual and textual features. Each cluster is considered a perceptual concept in the medianet.

Clustering is the process of discovering natural patterns in data by grouping similar data items [66]. We use a diverse set of clustering algorithms: the k-means algorithm, the Ward algorithm, the k-Nearest Neighbors algorithm (KNN), the Self-Organizing Map algorithm (SOM), and the Linear Vector Quantization algorithm (LVQ). The rationale for selecting each algorithm follows. The k-means algorithm is one of the simplest and fastest clustering algorithms. The Ward algorithm has been shown to outperform other hierarchical clustering methods. The k-nearest neighbor does not require any metric (i.e., only the nearest neighbors of each item to be clustered) and can avoid the globular biases of other clustering algorithms. SOM and LVQ are neural network clustering algorithms that are capable of learning feature weights. Besides, LVQ allows treating annotations as labels driving the clustering; this is one way to integrate text and images in the clustering.

We can cluster images based on any combination of visual and/or textual features. Regarding local visual features, we can cluster the images based on the features of all the regions, the largest regions, or the center region. We can also generate clusters based on any combination of textual and visual features by concatenating the corresponding feature vectors. The mean and the variance of the bins of concatenated feature vectors

are normalized to zero and one, respectively. Concatenation and normalization is the second way of integrating visual and textual features in the clustering. The dimension of feature vectors can be reduced using Latent Semantic Indexing (LSI) [42], which has the effect of uncorrelating feature vector bins.

4.3.3 Perceptual Relationship Extraction

SOM/LVQ and Ward clustering algorithms already provide similarity and hierarchical relationships among the clusters (given by the mesh and hierarchy of clusters), respectively. Additional relationships between clusters are discovered by analyzing the feature similarity and the statistical dependencies between the clusters.

Each cluster is said to have similarity relationships with its k nearest cluster neighbors. The distance between two clusters is calculated as the distance between the corresponding centroids. The number of neighbors could be set from 2 to 4 because that is the cluster neighbor range for SOM and LVQ clusters. We propose new methods for extracting relationships "equivalent", "specializes", "co-occurs", and "overlaps" between clusters based on cluster statistics, which are defined in Table 4.1. For example, if two clusters use the same features and their co-occurrence conditional probabilities are one or very close to one, they are considered equivalent. Such relationships will prove to be useful in summarizing medianets (see section 4.7.4).

Table 4.1: Rules for discovering statistical relationships between clusters where FD(c) represents the features used to generate cluster c, p(c1 | c2) is the probability of an image to belong to cluster c1 if it belongs to cluster c2, α is a positive real number smaller but close to one, and β is a positive real number smaller than α.

	FD (c1) = FD(c2)	**FD (c1) ≠ FD(c2)**
$P(c1|c2), p(c2|c1) > \alpha$	c1 equivalent to c2; and vice versa	c1 co-occurs with c2; and vice versa
$0 < p(c1|c2) < \beta; p(c2|c1) > \alpha$	c1 specializes c2; c2 generalizes c1	c2 co-occurs with c1
$0 < p(c1|c2), p(c2|c1) < \alpha$	c1 overlaps c2; and vice versa	--

4.4 Semantic Knowledge Discovery

The proposed approach for extracting semantic knowledge from annotated images, which have already been clustered as described in section 4.3, consists of three steps, as shown in Figure 4.3 [10][16]. First, words are tagged with their part-of-speech information, and annotations chunked into word phrases (e.g., noun and verb phrases). Then, semantic concepts are extracted by disambiguating the senses of the words with a novel method that uses both WordNet and the image clusters. Finally, semantic relationships and additional concepts relating the detected senses are found in WordNet. This section discusses each step in detail.

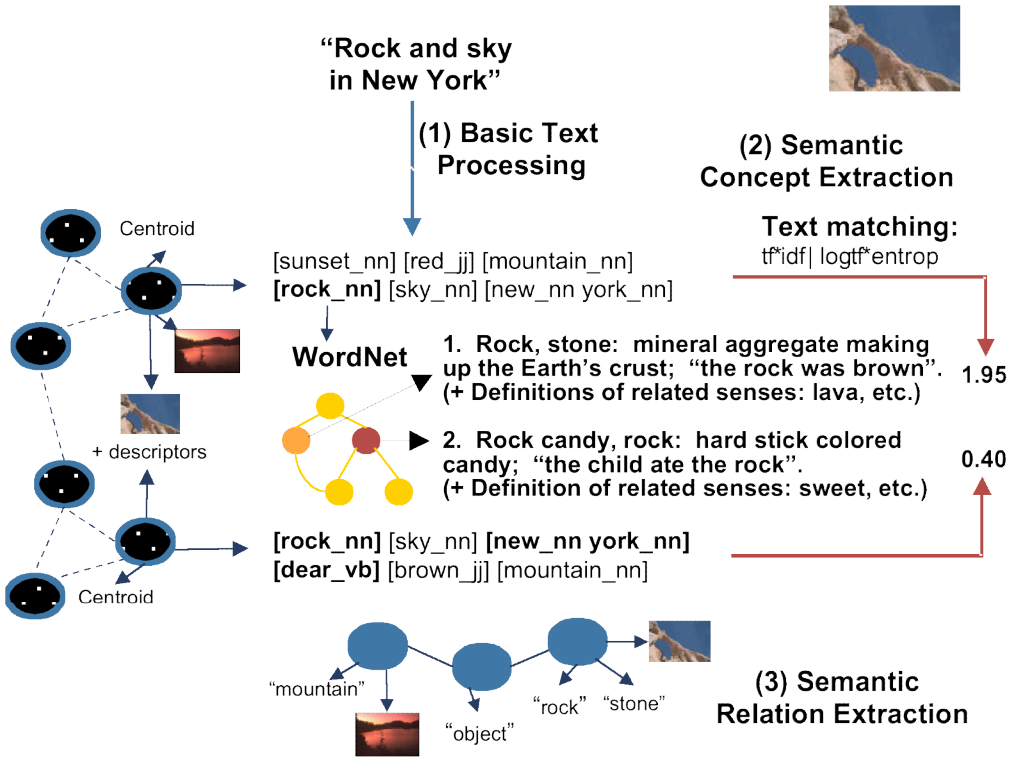


Figure 4.3: Semantic knowledge extraction process: (1) first, the textual annotations are tagged with their part-of-speech ("_nn", "_vb", "_jj" and "_rb" for nouns, verbs, adjective and adverbs, respectively) and chunked into word phrases; (2) then, semantic concepts (plain ellipses) are extracted by disambiguating the senses of the words using WordNet and the image clusters; and, (3) finally, semantic relationships (plain lines) among senses are found in WordNet.

4.4.1 Basic Text Processing

During this step, the words are stemmed and tagged with their part-of-speech, and stop words and non-content words are discarded, as described in section 4.3.1. Then, the textual annotations are chunked into noun and verb phrases (e.g., the sentence "I love New York" has two noun phrases "I" and "New York", and one verb phrase "love") [91]. In addition, single words are grouped into compound words (e.g., "New York" in Figure

4.3 is one compound word with one meaning). For the recognition of compound words, we detect noun and verb phrases containing only nouns or verbs, respectively. Then, different combinations of the words, starting from the ones with more words and preserving word ordering, are searched in WordNet. If a word search is successful, the words are removed from the following word combinations until all the combinations have been searched. As an example, the noun phrase "New York" in Figure 4.3 will generate the following word searches: "New York ", "New" and "York"; the first search is successful so no additional searches are executed.

4.4.2 Semantic Concept Extraction

The second step in the semantic knowledge extraction process is to disambiguate the senses of the remaining words using WordNet and the image clusters. Each detected sense is a semantic concept in the medianet.

The intuition behind the proposed approach is that the images that belong to the same image cluster are likely to share some common semantics. The common semantics although general can help the word-sense disambiguation (e.g., images of animals and flowers in vegetation have similar global color and share semantics such as "nature" and "vegetation"). The proposed technique also follows the two principles for word-sense disambiguation: consistent sense for a word and semantically relatedness of nearby words in the annotations of clusters. The word-sense disambiguation procedure consists of two basic steps (see Figure 4.3). First, the senses of words annotating the images in a

cluster are ranked based on all the annotations of (the images in) the cluster. The ranking method is explained below. An image can belong to several clusters; the second step is to add the ranks of the senses for the same word and image for the different clusters to which the image belongs. The more relevant the concept, the higher the rank. Therefore, the detected sense for a word is the highest ranked sense.

We rank the different senses of a word for an image in a cluster by matching the definitions of the possible senses listed by WordNet to the annotations of all the images in the cluster using word weighting schemes, i.e., tf*idf and log tf*entropy [46]. In this process, the definition of each sense is considered to be a document in a collection; and the query keywords, the annotations of all the images in the cluster. The definition of a sense (e.g., sense "rock, stone" in Figure 4.3) is constructed by concatenating the synonym set (e.g., "rock, stone"), the meaning (e.g., "mineral aggregate making up the Earth's crust"), and the usage examples of the sense (e.g., "the work was brown") together with the definitions of directly or indirectly related senses (e.g., sense "lava", which specializes sense "rock, stone") provided by WordNet. Different weights are assigned to the synonym set, the meaning, and the usage examples of a sense, and the definitions of related senses. As an example, higher weights should be assigned to the synonym set (e.g., 1.0 for "rock" and "stone") compared to the usage examples (e.g., 0.8 for "rock" and "brown"), and to the definition of a sense compared to the definition of related senses (e.g., 1.0 for definition of sense "rock, stone", 0.8 for definition of sense "lava").

4.4.3 Semantic Relationship Extraction

The third step is to discover semantic relationships among the semantic concepts. During this process, we find the paths connecting every pair of detected senses in WordNet, either directly or through intermediate senses. All the semantic relationships and the intermediate senses on these paths are added to the extracted semantic knowledge. Therefore, the constructed knowledge is not restricted to the detected senses. For example, in Figure 4.3, senses "mountain" and "rock, stone" are connected through the concept "object", their common ancestor, and specialization relationships between them. Table 3.1 includes the semantic relationships in WordNet together with definitions and examples.

4.5 Multimedia Knowledge Summarization

Our proposed approach for summarizing medianets, i.e., multimedia concept networks, consists of reducing the number of concepts and relationships in the network in three steps, as shown in Figure 4.4 [10][14]. First, the distances among the concepts in the network are calculated. Then, similar concepts are clustered together based on the concept distances. Finally, the summary is generated based on the concept clusters. This section discusses each step in detail. In a preliminary stage, the least frequent concepts can be discarded.

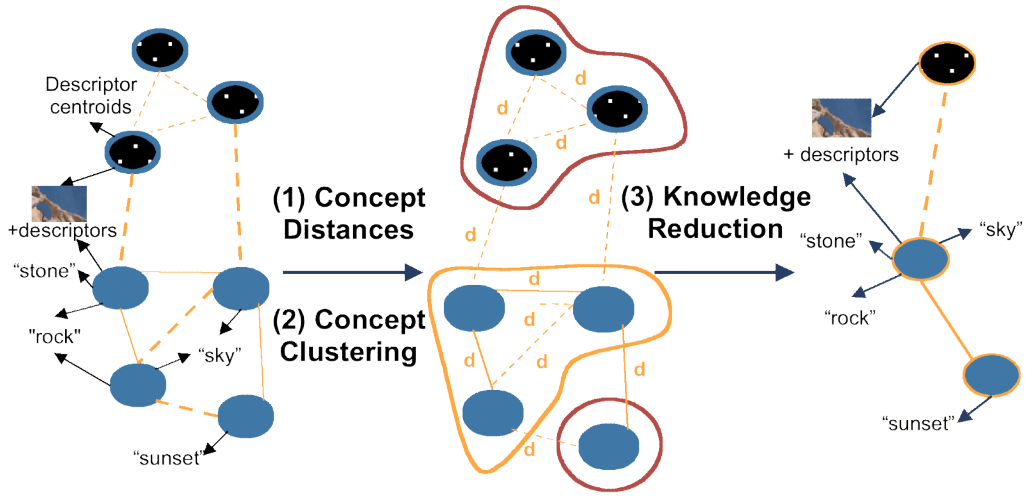


Figure 4.4: Multimedia knowledge summarization process: (1) first, distances between concepts are calculated; (2) then, similar concepts are clustered together; and, (3) finally, the summary is constructed based on the concept clusters.

4.5.1 Concept Distances

The first step in summarizing medianets is to calculate the distances between concepts using a novel technique based on concept statistics and network topology.

There are many proposed methods for calculating distance or similarity between concepts in semantic concept networks such as WordNet [68][142]. Some methods rely uniquely on hierarchical specialization/generalization relationships between concepts [68] whereas others take into account all the semantic relationships [142]. There are methods that use exclusively the concept network topology [142] while others combine both concept network topology and concept statistics in a corpus [68]. Techniques using concepts statistics and specialization/generalization relationships seem to outperform other approaches. As an example, Budanitsky and Hirst [27] evaluated five concept

distance measures using WordNet in a real-word spelling error correction system in which the concept distance measure proposed by Jiang and Conrath [68] was found to outperform the rest.

Jiang and Conrath's concept distance measure [68] takes into account the specialization/generalization hierarchy and the statistics of concepts for calculating concept distances. The distance of a relationship in the concept hierarchy is the information content, as defined in information theory, of the child concept given the parent concept, i.e., of encountering an instance of the child concept c given an instance of the parent concept c', as follows:

$$\mathrm{dist}_{JC}(c, c') = \mathrm{IC}(c \mid c') = -\log(p(c \mid c')) \tag{4.1}$$

where IC(x) is the information content of x, p(c) is the probability of encountering an instance of concept c, and $p(c \mid c')$ is the probability of encountering an instance of concept c given an instance of concept c'. Then, the distance between any two concepts c1 and c2 is obtained using the information content of the two concepts together with the most specific common ancestor (msca operator), as follows:

$$\mathrm{dist}_{JC}(c1, c2) = 2\ \log(p(\mathrm{msca}(c1, c2))) - (\log(p(c1)) + \log(p(c2))) \tag{4.2}$$

This distance measure is limited to a specialization/generalization hierarchy between concepts; it does not take into account any other relationships between concepts. This is

important because we consider medianets with arbitrary relationships that may not be organized around a specialization/generalization hierarchy. However, none of the concept distance measures that consider additional relationships take into account concept statistics.

We propose a novel concept distance measure that generalizes the measure proposed in [68] to an arbitrary concept network with different relationships between concepts and media examples. The proposed concept distance measure is based on concept statistics and network topology so it is not limited to specialization/generalization relationships. Instead, it supports relationships such as "contains", "entails", and "overlaps". First, we calculate the distance of each relationship in the medianet based on concept statistics. The distance between any two concepts is then obtained as the distance of the shortest distance path (which may consist of several hops) between the two concepts in the network. The distance of a relationship r connecting concepts c and c' is calculated as follows:

$$\begin{aligned} dist(c,c',r) &= p(c)\, IC(c',r \mid c) + p(c')\, IC(c,r \mid c') \\ &= -p(c)\, \log(p(c',r \mid c)) - p(c')\, \log(p(c,r \mid c')) \end{aligned} \tag{4.3}$$

where $p(c, r \mid c')$ is the probability of encountering an instance of concept c through relationship r given an instance of concept c'. We assume binary relationships, i.e., relationships that only have two vertices, a source and a target. The proposed concept distance satisfies the properties of a distance function.

The intuition behind equation (4.3) is the following: the distance of a relationship between two concepts increases with the concept probabilities but decreases with the conditional probabilities of the concepts for the relationship. In other words, the distance between concepts that are rare is lower than the distance between concepts that are common (e.g., concepts "tiger" and "leopard", which specialize concept "cat", are more related than concepts "object" and "organism", which specialize concept "entity"). This is similar to the process of "depth-relative scaling" [142], which states that sibling concepts deep in the hierarchy are more related than sibling concepts higher in the hierarchy. In addition, if two concepts have a higher conditional probability, they are more similar to each other. This is the same principle that governs Jiang and Conrath's concept distance [68]. This distance corresponds to the first information content term in equation (4.3), IC(c', r|c), when concept c is the parent node of concept c' in the specialization/generalization hierarchy.

There are different approaches for calculating concept statistics such as WordNet's senses in a text corpus. The approach [117], often used in conjunction with measure [68], obtains the frequency of a concept by adding the occurrences of specialized concepts to the strict occurrence of a concept. In a similar way, we first find strict concept frequencies for each concept, $freq_o(c)$, by summing up the number of times a concept is instantiated in each image. As an example, concept "house" would have a frequency of two for an image whose annotation contains the word and sense "house" twice. The inferred concept frequencies are then propagated in the medianet, e.g., an instance of

concept "dog" is also an instance of concept "animal". Considering a relationship r between concepts c and c', a different fraction of the frequency of concept c is added to the frequency of concept c' based on relationship r, and vice versa. In mathematical terms, the inferred frequency of concept c is calculated using the following system of equations:

$$\text{freq}(c) = \text{freq}'(c) + \sum_{r \in \text{relationsWithSrc}(c)} w_{t \to s}(r)\ \text{freq}(\text{tgt}(r)) + \sum_{r \in \text{relationsWithTgt}(c)} w_{s \to t}(r)\ \text{freq}(\text{src}(r)) \qquad (4.4)$$

where relationsWithSrc(c) and relationsWithTgt(c) are the sets of relationships that have c as source and target, respectively; src(r) and tgt(r) are the source and target nodes of relationship r; $w_{s\_t}(r)$ and $w_{t\_s}(r)$ are the relationship propagation weights for relationship r from source to target and from target to source, respectively; concepts(N) is the set of concepts in medianet N; and freq'(c) is proportional to freq_o (c).

The system specified by the recursive equation (4.4) represents the relationship of the final inferred concept frequencies. This system is solved using a simple, iterative method that consists of starting with any set of concept frequencies and iterating the computation until the system converges. The system can have a solution in spite of loops in the concept network if the relationship propagation weights are not too large. This method is similar to the one that the Google search engine uses to calculate PageRanks for rating web pages [112]. The relationships in the medianet affect the inferred concept frequencies and, therefore, the concept distances through the relationship propagation

weights, $w_{s\_t}(r)$ and $w_{t\_s}(r)$, which can be learned or specified by experts (see Table 4.2 for examples). Finally, the probability of concept c and the conditional probability of concept c through relationship r given concept c' are calculated as follows:

$$p(c) = \frac{freq(c)}{\sum_{c \in concepts(N)} freq_o(c)} \tag{4.5}$$

$$p(c, r \mid c') = \frac{w_{s \to t}(r)\, freq(src(r)) + w_{t \to s}(r)\, freq(tgt(r)) - w_{s \to t}(r)\, w_{t \to s}(r)\, freq(tgt(r) \cap src(r))}{freq(c')} \tag{4.6}$$

Table 4.2: Propagation weights for several semantic and perceptual relationships from source to target, $w_{s\_t}(r)$, and vice versa, $w_{t\_s}(r)$. These weights are used to calculate the concept frequencies.

Relation	Source to Target	Target to Source
Equivalent	1.0	1.0
Specializes	1.0	0.0
Contains	0.5	0.0
Causes	0.5	0.0

4.5.2 Concept Clustering

The second step in the multimedia knowledge summarization process is to cluster the concepts based on their distances. Similar concepts are grouped together in as many clusters as the desired number of concepts in the summarized medianet. The number of

clusters can be decided by a human supervisor or set automatically using the measures presented in the section 4.6 as described in section 4.7.4.

We can currently cluster the concepts in a medianet using a modified KNN clustering algorithm and the spectral clustering algorithm [107]. The KNN clustering algorithm was selected because of the continuity and the non-globular shape of the resulting clusters. Moreover, the KNN clustering algorithm does not require a specific distance function. The spectral clustering algorithm was selected for three reasons: it is directly applicable to clustering concepts in a medianet (it evolved from spectral graph clustering); the clusters are connected in the medianet; and the number of members in each cluster is quite balanced (i.e., clusters tend to have similar number of members). Algorithms that generate more balanced clusters are less sensitive to outliers and more appropriate for hierarchical clustering and browsing because the cluster hierarchy is shorter (see section 6.3).

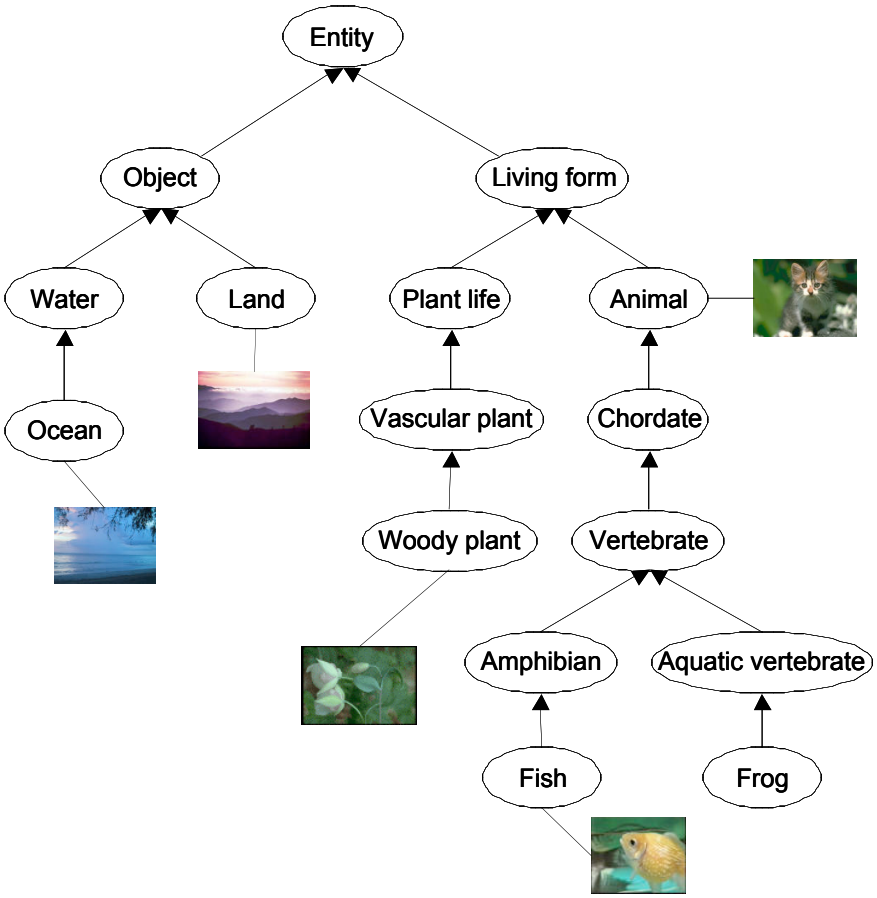
The modified KNN clustering algorithm groups concepts into a given number of clusters. Whereas the KNN clustering algorithm merges the clusters of two data items with at least k_t shared neighbors within k neighbors [67], the modified KNN clustering algorithm merges the clusters of the two data items with the largest number of shared neighbors until a given number of clusters is reached. The input to the clustering algorithm is the k nearest concepts of each concept and the desired number of clusters. We support different weighting schemes of shared neighbors [67].

4.5.3 Knowledge Reduction

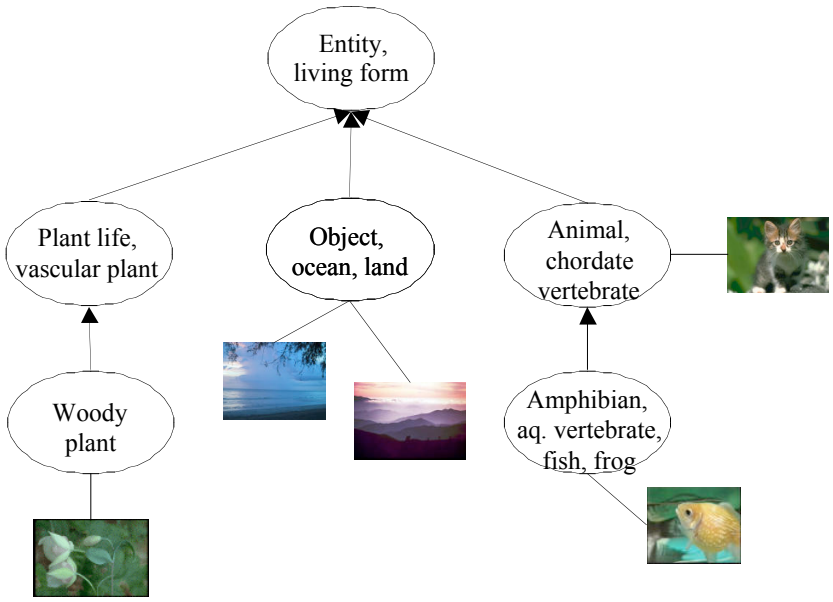
The medianet summary is generated using the concept clusters. Every concept in the medianet summary corresponds to a concept cluster, i.e., a group of concepts or super concept. A super concept in the medianet summary inherits the text and images of cluster members. If all the concepts in a concept cluster are semantic concepts, then the corresponding super concept in the medianet summary is a semantic concept; otherwise, it is a perceptual concept. The relationships among super concepts in the medianet summary are the relationships between the concepts in different concept clusters. In other words, we keep all the relationships among concepts in different clusters and assign them to the corresponding super concepts.

Figure 4.5.a and Figure 4.5.b show, respectively, a medianet of 16 concepts and a summary of the medianet with 6 super concepts. In visualizing medianets, decisions should be made for labeling and displaying the concepts and the relationships between them. For example, the labels of a super concept can be selected as the textual examples corresponding to the centroid or the most probable concept in the concept cluster. Regarding the relationships between two concepts, they could be represented using only the relationship with highest or lowest distance, among other strategies. In chapter 6, we propose and evaluate several of these methods for displaying and visualizing medianets. In Figure 4.5, the concept labels were manually chosen for illustration purposes. In general, super concepts may include both perceptual and semantic concepts and there

may be multiple relationships, both uni- and bi-directional, of different types between super concepts.



a)



b)

Figure 4.5: Examples of a) a medianet of 16 concepts and b) a summary of the medianet with 6 super concepts. The relationships are specialization relationships. The arrows indicate the direction of the relationship from source to target. For the medianet summary in figure b), the concept labels were manually chosen for illustration purposes. In general, super concepts may include both perceptual and semantic concepts and there may be multiple relationships, both uni- and bi-directional, of different types between them.

4.6 Multimedia Knowledge Evaluation

In this section, we propose automatic ways for measuring the consistency, completeness, and conciseness of medianets. These are three of the five criteria identified by Gomez-Perez [52] for expert evaluation and assessment of semantic ontologies. The other two criteria, expandability and sensitiveness (i.e., how new definitions or changes in existing definitions affect the properties on the ontology, respectively), are not considered because they usually depend on the ontology management. In addition, at this time, we

do not consider the problem of incrementally adding new images or annotations to the medianet. The goal of the proposed measures is to evaluate the goodness of a medianet in an automatic and application independent way.

4.6.1 Consistency

Consistency refers to whether it is possible to obtain contradictory conclusions from valid input definitions. In terms of concept distances, the consistency of medianets can be calculated based on the differences on the distances between two concepts through different paths in the medianet. The larger the distance spread between concepts, the more inconsistent or contradictory the different paths connecting the concepts.

We propose to measure the inconsistency of medianets by calculating the spread of the total distances of the k shortest distance paths between every pair of concepts with respect to the shortest distance path, as follows:

$$\mathrm{ICST}(N) = \log\left(\frac{\sum_{c,c' \in \mathrm{concepts}(N)} \sum_{i=1}^{i=k} (d(c,c',i) - d(c,c',1))^2}{|\mathrm{concepts}(N)|^2 \; k} + 1\right) \tag{4.7}$$

where $|\mathrm{concepts}(N)|$ is the number of concepts in medianet N, k is the number of shortest distance paths considered between any two concepts, and d(c,c',i) is the distance between concepts c and c' through path i. The k shortest distance paths are ordered from

shortest to longest distance starting at i=1 up to i=k. The lower ICST(N), the more consistent the medianet.

4.6.2 Completeness

Completeness refers to the completeness of both the ontology and the definitions in the ontology. The only way for measuring the completeness of medianets would be the direct comparison with target medianets, which are rarely available. Instead, we measure the information (randomness), the graph completeness, and the concept-category correlation of medianets. The more informative (random), the more complete the graph, and the higher correlation with category labels; the more complete the medianet.

We propose to measure the randomness of information in a medianet by calculating the concept entropy, as follows:

$$CPT\_H(N) = -\sum_{c \in concepts(N)} p(c) \log(p(c)) \tag{4.8}$$

where p(c) is the probability of concept c obtained as described in section 4.5.1. We measure the graph completeness of a medianet by adapting the formula of graph density (i.e., ratio of the number of relationships in the graph by the maximum number of possible relationships) to weighted relationships, as follows:

$$CPT\_D(N) = \frac{\sum_{r \in relations(N)} [1 - d(r)/dmax]}{|concepts(N)|(|concepts(N)|-1)} \tag{4.9}$$

where relations(N) is the relationships in the medianet N, d(r) is the distance of relationship r, and d_{max} is the maximum distance for a relationship. The higher CPT_H(N) and CPT_D(N), the more complete the medianet.

If category labels are available for the images, we propose an entropy-based criterion to evaluate the completeness or correlation of the concepts with the category labels. If L $\{l_1, \ldots l_m\}$ and C = concepts(N) = $\{c_1, \ldots c_n\}$ are the sets of category labels and concepts, respectively, the correlation of concepts C with respect to category labels L can be calculated as the harmonic mean of one minus the mean entropies of the categories within each concept (INH(C)), and of each category over the concepts (INH(L)), normalized from 0 to 1, as follows:

$$CPT\_CH(C, L) = \frac{2\, INH(C)\, INH(L)}{INH(C) + INH(L)} \tag{4.10}$$

where

$$\mathrm{INH(C)} = 1 - \sum_{j=1}^{n} p(c_j) \sum_{i=1}^{m} p(l_i \mid c_j) \log(p(l_i \mid c_j)) \Big/ \log(m)$$
$$\mathrm{INH(L)} = 1 - \sum_{i=1}^{m} p(l_i) \sum_{j=1}^{n} p(c_j \mid l_i) \log(p(c_j \mid l_i)) \Big/ \log(n) \qquad (4.11)$$

The harmonic mean ensures that CPT_CH(C,L) is close to one only if both INH(C) and INH(L) are close to one. The closer PCT_CH(C,L) to one, the better the medianet N fits the labels L. PCT_CH(C,L) is equal to one if and only if the number of the concepts and the categories are the same and the images in each concept are exactly the same as the images in a category.

4.6.3 Conciseness

Conciseness refers to whether all the information in the ontology is precise, necessary, and useful. Using concepts distances, we calculate the redundancy of a medianet as the number of null eigen values in the concept distance matrix. The larger the number of null eigen values, the more redundant (i.e., less concise) the medianet.

Our proposed way to evaluate the conciseness of a medianet N is by comparing the number of concepts and the rank of the concept distance matrix, as follows:

$$\mathrm{ICCS(N)} = \frac{|\,\mathrm{concepts(N)}\,| - \mathrm{rank(M)}}{|\,\mathrm{concepts(N)}\,|} \qquad (4.12)$$

where M is the concept distance matrix and rank(M) is the rank of the matrix M. The elements of the matrix M are the pair-wise distances between every pair of concepts. The lower ICCS(N), the more concise the medianet.

4.7 Evaluation Experiments

In this section, we describe the test set together with the experiments performed for evaluating the proposed techniques for perceptual and semantic knowledge discovery, and for multimedia knowledge summarization.

Semantic and perceptual multimedia knowledge in the form of a medianet was extracted from a collection of 3,624 annotated images as described in sections 4.3 and 4.4, respectively. Concept-category correlation (CPT_CH(C,L)) and the word-sense disambiguation accuracy were used to evaluate the resulting image clusters and word senses, respectively.

The medianet extracted from a smaller collection of 271 images was summarized in different sizes using the modified KNN clustering algorithms as described in section 4.5. Distance spread (ICST(N)), concept entropy (CPT_H(N)), graph density (CPT_D(N)), and concept redundancy (ICCS(N)) were used to compare the extracted and the summarized medianets generated using the proposed techniques with respect to several baseline approaches.

4.7.1 Test Set

The test set was a diverse collection of 3,624 nature and news images from Berkeley's CalPhotos collection (http://elib.cs.berkeley.edu/photos/) and the ClariNet news newsgroups (http://www.clari.net/), respectively. The images in CalPhotos were already labeled as plants (857), animals (818), landscapes (660), or people (371). The news images from ClariNet were categorized into struggle (236), politics (257), disaster (174), crime (84), and other (67) by researchers at Columbia University. The nature and news images had annotations in the form of keywords and sentences, respectively (see Figure 4.6).

Figure 4.6: Example of a news image (left) and a nature image (right) with their textual annotations.

4.7.2 Perceptual Knowledge Discovery

In this section, we present the setup and discuss the results of the experiments performed for evaluating the proposed techniques for perceptual knowledge discovery.

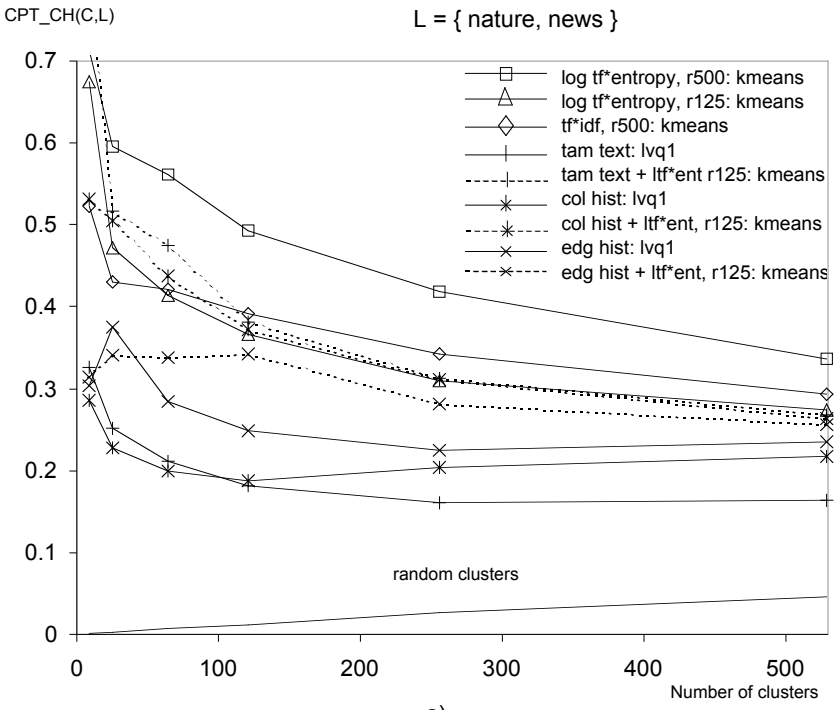
4.7.2.1 Setup

During the perceptual knowledge extraction process, the images were scaled down to a maximum height and width of 100 pixels and segmented to at most 16 regions. Words that appeared less than 5 times in the images annotations were discarded for the

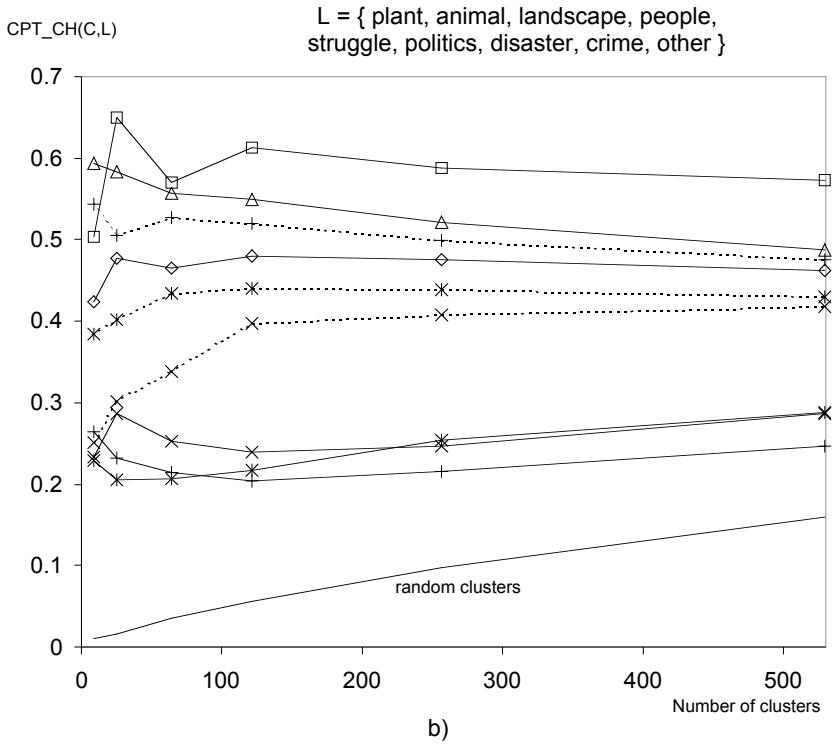
extraction of the textual features, whose dimensionality was further reduced to 500 and 125 using Latent Semantic Indexing (LSI). Clustering was done using different algorithms -k-means, SOM, LVQ, and KNN-, different features -color histogram, Tamura texture, edge direction histogram, the features of the 16 regions, the largest region's features, the center region's features, the tf*idf feature, and the log tf*entropy feature; and different number of clusters - ranging from 9 to 529. The SOM and LVQ maps were made square. The labels used for LVQ clustering algorithms were the category labels listed above.

4.7.2.2 Results

The criterion used to evaluate the image clusters generated during the perceptual knowledge extraction process was the concept-category correlation, CPT_CH(C,L), using the primary categories {nature, news} and the secondary categories {plant, animal, landscape, people, struggle, politics, disaster, crime, other} as the labels L. Figure 4.7.a and Figure 4.7.b show the results obtained for the best clustering algorithm for different features for the two category sets, respectively. The results for the local (region) visual features were excluded from the figure because they were the worst due, likely, to the use of the Euclidean metric in building the clusters instead of specialized metrics [171].


CPT_CH(C,L)
L = { nature, news }
log tf*entropy, r500: kmeans
log tf*entropy, r125: kmeans
tf*idf, r500: kmeans
tam text: lvq1
tam text + ltf*ent r125: kmeans
col hist: lvq1
col hist + ltf*ent, r125: kmeans
edg hist: lvq1
edg hist + ltf*ent, r125: kmeans
random clusters
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
100
200
300
400
500
Number of clusters

a)


CPT_CH(C,L)
L = { plant, animal, landscape, people,
struggle, politics, disaster, crime, other }
random clusters
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
100
200
300
400
500
Number of clusters

b)

Figure 4.7: Concept-category correlation (CPT_CH(C,L)) results (y axis) per number of clusters (x axis) for (a) the primary categories and (b) the secondary categories. "col hist" is color histogram, "tam text" is Tamura texture, "edg hist" is edge direction histogram, "ltf*ent" is log tf*entropy, "lvq1" is LVQ with primary categories, and "r125/500" is LSI for a reduced dimensionality of 125/500 bins. Results for random assignment into clusters are provided for baseline comparison (i.e., "random clusters").

Figure 4.7 also displays results in concatenating the 125-bin log tf*entropy feature and each visual feature with bin normalization but no LSI. The results with normalization and LSI were very similar that together with the small number of null eigen values for concatenated textual-visual feature vectors (e.g., 3 bins for 166-bin color histogram + 125-bin log tf * entropy) shows the high independence of visual and textual features. The figure also includes results for randomly assigning images into clusters for baseline comparison (i.e., "random clusters" in Figure 4.7).

4.7.2.3 Discussion

As can be seen in Figure 4.7, both textual and visual features enable the discovery of useful knowledge because their results are well above random behavior. As expected textual features are more powerful than visual features and log tf*entropy feature outperforms tf*idf feature. Some concatenated textual-visual features slightly outperform the individual textual feature for the primary categories but not for the secondary categories probably because the latter categories are less visually separable. This indicates that both kinds of features should be integrated in the knowledge extraction process in providing different kinds of useful knowledge. Please, note that the

results for integrating 500-bin textual feature with visual features are not shown in the figure for clarity; however, their tendency is similar to the integrated 125-bin textual-visual features. For example, for 64 clusters, the integrated 500-bin textual + Tamura texture feature has also greater than the value for the 500-bin textual feature for the primary categories (see Figure 4.8). Although not shown in the results, the trend of INH(C) and INH(L) is monotonically increasing and decreasing, respectively.

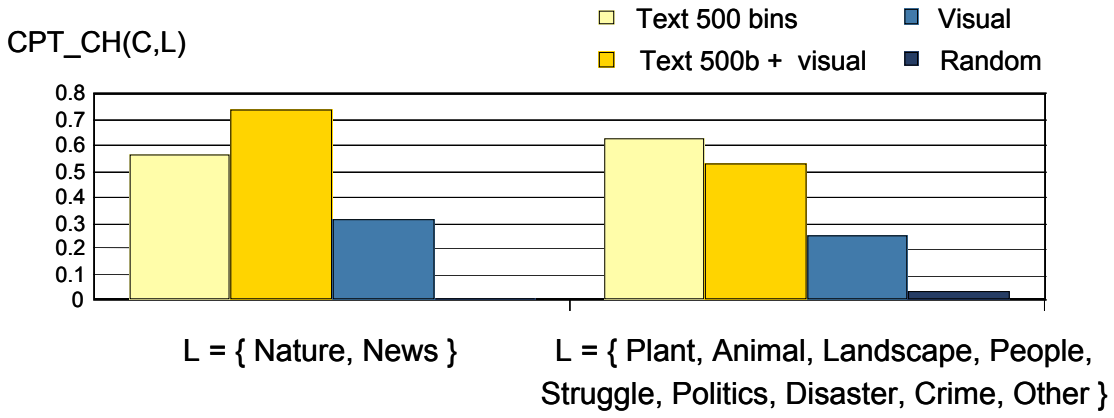


Figure 4.8: Concept-category correlation (CPT_CH(C,L)) results for 64 clusters based on the best textual feature of 500 bins (Text 500 bins), visual feature (Visual), and combination of the two (Text 500b + visual) for the primary and the secondary categories. Results for random assignment into clusters are provided for baseline comparison.

4.7.3 Semantic Knowledge Discovery

In this section, we present the setup and discuss the results of the experiments performed for evaluating the proposed techniques for semantic knowledge discovery.

4.7.3.1 Setup

During the semantic knowledge extraction process, the sense definitions were generated for clustering the images using different combination of features into different numbers of clusters. In addition, we experiment assigning different weights to the synonym set with respect to the meaning and usage examples of a sense, and to the definitions of directly and indirectly related senses. Lof tf * entropy was used to match sense definitions and cluster annotations using the cosine metric.

4.7.3.2 Results

The criterion to evaluate the word-sense disambiguation process was the word-sense disambiguation accuracy, in other words, the percentage of words correctly disambiguated. We manually generated the ground truth for the annotations of 10% of randomly selected images in the collection; no training was needed.

Table 4.3 shows the results for the clusters (with some features and number of clusters) that provided the highest and lowest word-sense disambiguation accuracy. We refer to these as the best image clusters (BI) and the worst images clusters (WI). For baseline comparison, Table 4.3 also includes the results for considering each image in its own cluster using the proposed techniques for WSD in section 4.4.2 (TT, WSD using only text), for selecting the most frequent sense of each word (MF, WordNet returns the senses for each word from most to least frequent), and for randomly picking between the

possible senses of each word (RD). The accuracy results are separated for the nature and the news images, and for nouns, verbs, adjectives, adverbs, and all the content words.

Table 4.3: Word-sense disambiguation accuracy (in percentages) for best image clusters (BI), worst image clusters (WI), image-per-cluster (TT), most frequent senses (MF), and random senses (RD). The results are provided separately for nature and news images, and for nouns, verbs, adjectives, adverbs, and all words. Column % indicates word percentages. The results for all the words are highlighted in italics.

	Nature Images					
	%	**BI** Best image clusters	**WI** Worst image clusters	**TT** Image per cluster	**MF** Most frequent senses	**RD** Random senses
Nouns	93.0	91.75	87.94	82.92	85.92	74.44
Verbs	1.08	66.67	40.74	59.26	44.44	44.44
Adjectives	5.72	58.04	41.43	40.85	55.71	44.29
Adverbs	0.20	100.0	33.33	37.50	100.0	75.00
All words	*100*	*89.20*	*85.29*	*84.72*	*83.80*	*72.42*
	News Images					
	%	**BI** Best image clusters	**WI** Worst image clusters	**TT** Image per cluster	**MF** Most frequent senses	**RD** Random senses
Nouns	60.8	66.06	57.07	57.88	68.59	45.86
Verbs	25.0	48.16	36.61	39.07	58.48	24.08
Adjectives	12.3	71.00	56.50	54.68	72.00	46.77
Adverbs	1.90	80.65	50.00	62.50	74.19	45.16
All words	*100*	*59.95*	*52.46*	*52.33*	*66.58*	*40.52*

4.7.3.3 Discussion

As shown in Table 4.3, for both image sets, best image clusters consistently outperforms cluster-per-image and random senses. For nature images, best image clusters and, often, worst image clusters provide better results than most frequent senses. The results for the news images are quite different: most frequent sense outperforms even best image clusters except for adjectives and adverbs. Several factors can explain the result differences between nature and news images: (1) WordNet has a more comprehensive coverage of nature concepts because several animal and plant thesauri were used in its construction; (2) the textual annotations of news images are well-formed phrases so there are more words that can potentially confuse the word-sense disambiguation process; (3) news images are more diverse and, therefore, their clusters may not be as "meaningful"; and, last but not least, (4) the gap between concepts and the visual features for news images is larger. The proposed approach for word-sense disambiguation is better than frequent sense for images with short annotations (e.g., one word "plant") that are clustered with more-extensively annotated and semantically related images (e.g., image annotated with "plant, flower, buttercup"); an example is shown in Figure 4.9.

Although not shown in Table 4.3, the best word-sense disambiguation results were obtained for 9 to 25 clusters, which reinforces the fact that visual clusters are useful for word-sense disambiguation. The use of different visual features or clustering algorithms had no obvious impact in the results. The only exception being that log tf * entropy, in some instances, concatenated with visual features (i.e., color histogram) consistently

provided good or the best clusters for word-sense disambiguation. Therefore, annotations of images with similar words and, often, similar features help the word-sense disambiguation. The best features could be selected for an image collection in a semi-supervised way if the correct senses for the annotations of a subset of the images are available (i.e., validation set). This process would be similar to feature selection for building image classifiers [63]. For generating the sense definitions, a reduction factor of 0.8 in the weights of the meaning and usage examples with respect to the synonym set, and of the definition of each relationship between the original sense and the related sense provided good results.

Plant
Yosemite National Park

Plant, flower, buttercup
Kings Canon National Park

Figure 4.9: Example where proposed method correctly disambiguates the sense of word "plant" but most frequent sense fails.

4.7.4 Multimedia Knowledge Summarization

In this section, we present the setup and discuss the results of the experiments performed for evaluating the proposed techniques for multimedia knowledge summarization.

4.7.4.1 Setup

A medianet was constructed for 271 randomly chosen images from the nature collection. A reduced set of images was used in these experiments to make the evaluation of the results tractable. We clustered the images based on color histogram, log tf * entropy, and a concatenated color histogram + log tf * entropy feature vector into 16 clusters for each feature. The nouns in the annotation field "what" of the images were the only ones used in the semantic knowledge extraction. The initial medianet had 790 semantic concepts, 48 perceptual concepts, 842 specialization relationships, 414 containment relationships, and 9 association relationships. Summaries of different sizes were generated from the initial medianet (including both semantic and perceptual concepts) using the modified KNN clustering algorithm. Concept occurrences were propagated based on the propagation relationship weights shown in Table 4.2, among others.

4.7.4.2 Results

Table 4.4 shows the values for distance spread (ICST(N)), concept entropy (CPT_H(N)), graph density (CPT_D(N)), and concept redundancy (ICCS(N)) obtained in the experiments evaluating the proposed techniques for summarizing medianets.

The first row of Table 4.4 shows the results for the medianet extracted from the image collection using the proposed concept distance (dist, see equation (4.3)) and the semantic distance proposed in [68] (dist $_{JC}$). The second row shows the results for a random version of the extracted medianet. The randomized medianet was generated by randomly

shuffling the images and concepts in the medianet. Table 4.4 also shows the results in summarizing the extracted medianet into summaries of 2, 4, 8, 16, 32, 64, 128, 256, and 512 concepts using our proposed concept distance and the semantic distance in [68].

Table 4.4: Distance spread (inconsistency, ICST), concept entropy (completeness, CPT_H), graph density (completeness, CPT_D), and concept redundancy (inconciseness, ICCS) results for the extracted medianet, a randomized medianet, and medianet summaries of 2, 4, 8, 16, 32, 64, 128, and 512 concepts using the proposed concept distance, dist (see equation (4.3)), and the distance in [68], $dist_{JC}$. The results for the extracted knowledge are highlighted in italics.

Knowledge	ICST Distance spread		CPT_H Concept entropy		CPT_D Graph density		ICCS Concept redund.	
	dist	**$dist_{JC}$**	**dist**	**$dist_{JC}$**	**dist**	**$dist_{JC}$**	**dist**	**$dist_{JC}$**
Extracted	*0.002*	*0.156*	*24.583*	*14.559*	*0.002*	*0.001*	*0.288*	*0.287*
Randomized	7.027	9.000	49.561	37.759	0.002	0.001	0.045	0.043
Summary 2	0.000	0.000	0.074	0.074	0.500	0.500	0.000	0.000
Summary 4	3.824	3.824	0.086	0.086	0.250	0.250	0.000	0.000
Summary 8	4.713	4.713	0.105	0.105	0.125	0.125	0.000	0.000
Summary 16	4.312	4.312	0.618	0.618	0.071	0.071	0.000	0.000
Summary 32	0.187	0.699	5.212	4.300	0.229	0.220	0.125	0.000
Summary 64	0.079	0.836	7.619	5.533	0.089	0.068	0.000	0.000
Summary 128	0.008	0.556	10.681	6.884	0.030	0.018	0.008	0.008
Summary 256	0.020	0.523	14.503	8.269	0.010	0.005	0.004	0.000
Summary 512	0.003	0.379	18.083	9.727	0.004	0.002	0.008	0.008

Table 4.5: Occurrence probabilities (in percentages) of the most frequent words in annotations and concepts in the summary of 32 concepts.

Words		**Summary of 32 super-concepts**	
Plant	5.78	Entity, living entity	100.0
Animal	5,51	Plant life, vascular plant	33.92
Flower	4.91	Biological group	32.60
Landscape	4.44	Family (biology)	29.88
Habitat	4.43	Woody plant	29.88
People	2.49	Genus (biology)	29.77
Bird	1.88	Bush, flower, herb, grass	29.77
Culture	1.88	Landscape, ocean, land	29.45
Chordata	1.21	Animal, vertebrate	22.07
Mammal	1.14	Amphibian, fish, frog	17.69

Table 4.5 lists the most frequent words in the annotations and concepts in the summary of 32 concepts. Figure 4.10 shows part of the medianet summary of 32 concepts, the sub-network that includes the most frequent concepts. The original medianet for the concepts in the top half of Figure 4.10 is depicted in Figure 4.5.

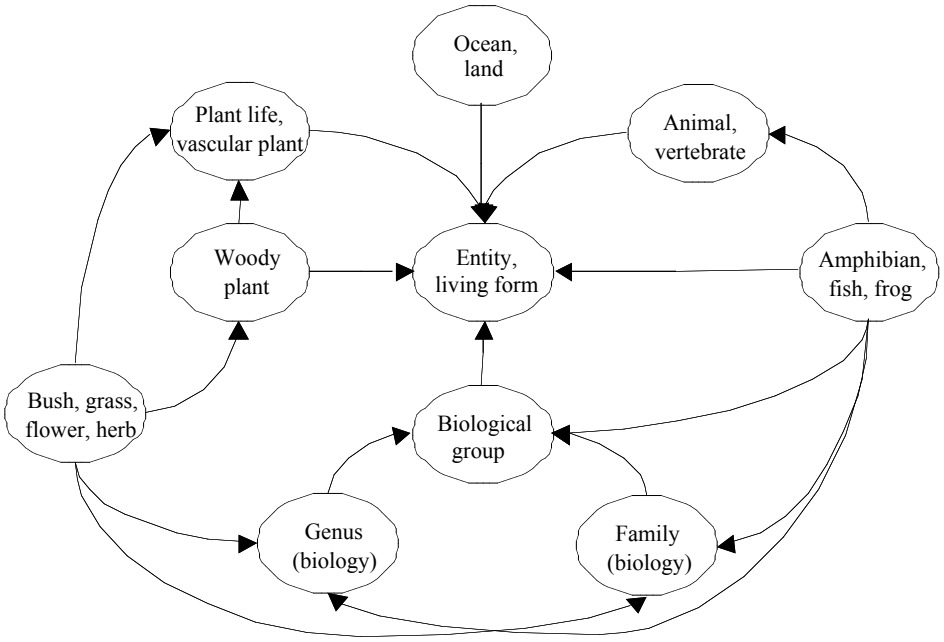


Figure 4.10: Sub-network of the medianet summary of 32 concepts. The relationships are specialization relationships. The arrows indicate the direction of the relationship from source to target. The concept labels were manually chosen for illustration purposes. In addition, all the relationships of the same type between two concepts were represented with only one arc in the figure. In general, super concepts may include both perceptual and semantic concepts and there may be multiple relationships, both uni- and bi-directional, of different types between them. The original medianet for the concepts in the top half of this figure is depicted in Figure 4.5.

4.7.4.3 Discussion

As expected, Table 4.4 shows that the randomized medianet has much higher concept entropy and lower concept redundancy, but also considerable larger distance spread. The graph density for the randomized medianet remains constant because the number of relationships does not change. Summarizing medianets increases the graph density and the distance spread, and decreases the concept entropy and the concept redundancy. In

the summarization process, similar concepts are grouped together so the resulting concepts are more different from each other, which causes the distance spread to increase The sharp concept entropy increase (and distance spread decrease) indicates the considerable difference between the summaries of 16 and 32 concepts. In fact, the summary of 16 concepts is composed of concepts that occur either very frequently or very rarely in the images. On the contrary, the summary of 32 concepts has a more uniform distribution of concept occurrences. We can then conclude that the proposed knowledge evaluation measures are useful for distinguishing between extracted and randomized knowledge, and for estimating the quality of knowledge. For example, these knowledge evaluation measures can be used to decide the number of concepts in which to summarize a medianet.

The use of different concept distances seems to have a considerable impact in the quality of the resulting summaries in spite of the large number of specialization relationships in the extracted medianet. The results for summaries of just a few concepts are almost identical for either of the concept distance measures. However, the proposed concept distance results in considerably higher concept entropy and lower distance spread, especially for summaries of 32 concepts and more, among others. Our proposed distance measure is therefore more significant and clear in comparing and observing the knowledge quality of randomized and summarized medianets, among others.

4.8 Summary

This chapter has presented novel methods for automatically discovering, summarizing, and evaluating multimedia knowledge from annotated images in the form of image clusters, word senses, and relationships, among them. These are essential for applications to intelligently, efficiently, and coherently deal with multimedia.

The proposed methods include (1) new techniques for discovering statistical and similarity relationships among image clusters (perceptual relationships and concepts); (2) a novel technique for disambiguating the senses of words and their relationships in image annotations that uses not only the annotations and WordNet but also the image clusters (semantic concepts and relationships); (3) a new technique for calculating distances between concepts, which is used to cluster concepts for summarizing medianets; and (4) automatic ways of measuring the consistency, completeness, and conciseness of medianets using notions from information and graph theory.

Experiments have shown both visual and textual features are useful in extracting different perceptual knowledge from annotated images; therefore, the integration of both kinds of features has potential to improve performance compared to individual features. The image clusters based on both visual and textual features have been shown to have a higher correlation with some of the considered categories than image clusters based on either visual or textual features (about 8% gain for 64 clusters). The evaluation of the proposed word-sense disambiguation approach has shown that using perceptual

knowledge in the form of image clusters can improve performance compared to most frequent senses and text-based word-sense disambiguation (about 6% gain for nature images). Additional experiments have shown the importance of good concept distance measures, as the proposed one, for clustering and summarizing medianets, and the potential of the proposed automatic measures for measuring the quality of medianets.

Large amounts of current multimedia are in the form of audio-visual data with optional caption and subtitle channels. For example, CNN receives over 300 hours of raw footage every day. In addition, there are many external resources of useful knowledge that are underutilized in multimedia information systems such as the Internet itself and the Cyc ontology. In the future, we plan to expand the multimedia knowledge discovery work in two directions: a) developing algorithms for other media such as moving pictures and audio, and b) integrating knowledge from other external resources apart from WordNet. As most multimedia collections are dynamic and grow with time, future work will also consist of proposing methods for modifying and updating the extracted medianets when images are added or removed from the collection. In addition, new procedures are needed to evaluate the quality of the knowledge represented by a medianets, i.e., specific instances of the MediaNet framework.

5 Knowledge-Based Image Classification

5.1 Introduction

This chapter focuses on novel methods for building classifiers for images based on medianets constructed from annotated images using WordNet. Current approaches in image classification lack flexibility: they are often constrained to specific domains or trained on limited data sets. In chapter 4, we presented our techniques for automatically discovering and summarizing medianets from annotated images in the form of image clusters, words senses, and relationships, among them.

In recent years, there has been a major increase in available multimedia and in technologies to access the multimedia. Users often want to retrieve, filter, and navigate multimedia at the semantic level (e.g., people and places depicted in the multimedia). However, current multimedia applications use features at the perceptual level (e.g., color, texture and cepstra coefficients) failing to meet user needs. For example, the study presented in [73] found that less than 20% of the attributes used by humans in describing images for retrieval were related to visual features. In addition, subject

hierarchy browsing was the most popular user operation in the web image search engine WebSEEk [136].

This chapter focuses on image classification. Image classifiers can be used to annotate images with semantic labels such as "person", "mountain", and "outdoors". However, current approaches lack flexibility: they are often constrained to specific domains or trained on limited data sets. As an example, current classifiers are often fine-tuned and limited to label images with a predefined set of semantic classes (e.g., "indoor" and "outdoor"). The semantic classes are usually manually specified by experts for each domain so they might not be applicable or relevant for other domains. In addition, prior work does not fully exploit existing image annotations and external knowledge resources such as WordNet for dynamically adapting and enhancing image classification.

In this chapter, we present novel approaches for building image classifiers from medianets discovered from annotated images using WordNet [12]. The main contributions of this work are the automatic selection of salient classes, and the combination of multiple classifiers, based on medianets discovered from annotated images. In addition, this work analyses the role of visual and textual features in image classification. As textual or joint visual-textual features perform better than visual features [111], we try to predict textual features from visual features for images without annotations. We use the term "MediaNet classifier" to refer to our image classification framework.

We propose to build a MediaNet classifier for a medianet in two steps. First, we train a meta-classifier to detect each concept in images using visual and textual features. A meta-classifier can be the result of combining several classifiers of different types or feature inputs. Then, a Bayesian network is learned using the meta-classifiers and the concept network. The presence of concepts in a new image is first detected using the meta-classifiers and this initial prediction is refined using Bayesian inference. Textual features are predicted for images without annotations using clustering and statistical approaches based on visual features extracted from images. There is usually more interest in detecting semantic concepts in images rather than perceptual concepts. In this case, perceptual concepts and relationships can be removed from the medianet before constructing the MediaNet classifier.

We have extensively evaluated the techniques proposed in this chapter using two different collections of images with annotations, a nature and a travel image collection. Experiments have shown that combining classifiers based on extracted and summarized medianets from annotated images using WordNet can result in superior accuracy (up to 15-30%) to individual classifiers and purely statistically learned classifier structures. As expected, textual and joint visual-textual features perform significantly better in classifying images than visual features. Predicting the textual features from visual features and using the joint visual-predicted textual features for image classification does not consistently improve the classification accuracy over using only visual features.

5.1.1 Outline of the Chapter

The rest of the chapter is organized as follows. In section 5.2, we review some related work on automatic image classification and annotation. In section 5.3, we present the construction of MediaNet classifiers for a collection of annotated images. We explain the way concepts are detected in new images in section 5.4. In section 5.5, we present the experiments performed for evaluating the proposed image classification methods. Finally, we conclude with a summary of the chapter and a discussion of future work in section 5.6.

5.2 Related Work

There have been many attempts to annotate and classify images and regions using semantic labels such as "indoor" and "face". Prior work on image annotation and classification can be categorized in terms of input features, classifier structure, and class selection. Many methods rely uniquely on perceptual features such as color and edge direction histograms [63][100][150][143][154][168]; whereas some also consider textual features from annotations or captions of the images [5][47][49][104][105][110][111]. There are some approaches that only use individual classifiers or joint distributions [5][47][100][168]; while others combine multiple classifiers for improved accuracy [49][63] [104][105][110][111][150][143][154]. Experts manually handpick the classes in many of these methods [49][104][105][110][111][150][143][154][168] to which the classifiers are often fine-tuned. However, there are frameworks where "knowledgeable"

users define their own classes and the relationships between the classes [63], and approaches that associate words in image annotations to new images or image regions [5][47][100]. A more detailed review of these works is presented below.

Early work on image classification trained simple classifiers on single visual features to label images with predefined semantic classes. An example is [168], which uses nearest neighbor and support vector machine classifiers to label images as indoor and outdoor scenes based on dominant orientation and color information. The former classifier is found to be better with color, the latter with dominant orientation. Szummer and Picard [143] demonstrate a performance gain in the indoor vs. outdoor classification problem by computing color and texture features on image blocks, classifying these blocks, and, then, combining the results to label the entire image. Different classifiers are evaluated including k-nearest neighbors, one-layer neural network, and mixture of experts, of which k-nearest neighbor is shown to outperform the rest. Vailaya et al. [154] also propose to use the k-nearest neighbor classifier to classify images into city vs. landscape, and sunset/sunrise vs. mountain vs. vegetation. The discriminative power of different color, texture, and edge features is measured using intra-class and inter-class similarities. This work also performs some basic classifier combination by dividing the sunset/sunrise vs. mountain vs. vegetation classification problem into two-class problems, sunset/sunrise vs. mountain/vegetation, and mountain vs. vegetation for the mountain/vegetation class. All these systems fine-tune the classifiers to limited data sets where they report classification accuracies of about or above 90%.

More recent systems have pursued more sophisticated ways for combining multiple classifiers and features based on manual or statistical models to boost the classification accuracy. Forsyth and Fleck [49] use body plans, which describe the visual aspect and geometry of body parts, for recognizing naked humans and horses. Specialized classifiers are trained for individual body parts and groupings, and are combined into a hierarchical classifier. Body plans are either built automatically using statistical learning techniques or specified manually by experts. In the system [104][105], a network of probabilistic multimedia objects, referred to as a multinet of multijets, serves as model for detecting objects and scenes in videos. Probabilistic models based on Gaussian Mixture Models and Hidden Markov Models are built for concepts such as rocks, sky, snow, water-body, forestry/greenery, and outdoor. Bayesian networks and factor graphs are used to combine the probabilistic models resulting in improved classification accuracy in [104] and [105], respectively. The topology of the Bayesian network is manually specified by the authors; whereas the parameters of the factor graphs are learned automatically. Similarly, Paek et al. [110][111] demonstrate an improvement in the classification accuracy by combining not only multiple classifiers but also a few semantic classes for images (e.g., "indoor"/"outdoor", "people"/"no people", and "city"/"landscape") using a Bayesian network. Example classification problems considered in this work are indoor vs. outdoor, sky vs. no sky, and vegetation vs. no vegetation. Manually constructed Bayesian networks consistently seem to perform better than statistically learned Bayesian networks. In Jaimes and Chang [63], users provide hierarchical models of object/scene and label training examples so users require a certain expertise. The system then learns

classifiers for each component of the hierarchical model by automatically selecting the best classification algorithms and features for each component. The final structure object/scene visual detector consists of a hierarchy of these classifiers as specified by the user in the model. A more general approach for fusing the output of individual classifiers is the technique proposed in [150]. Classifiers of different types using different features and parameters are fused by normalizing the confidence scores, aggregating their outputs using a combiner function, and selecting the optimal combination based on a validation set. This work uses more than 100 classes, which were also hand picked by experts in a community effort. Similar works that also assigned concepts from medium or large control vocabularies are [28][151].

In the prior work discussed above, the classes and structure of the classification frameworks were either decided by experts or specified by users. More general approaches attempt to label new images and, sometimes, their regions, with words in annotations of training images. The approach presented in [100] is based on co-occurrence statistics between words and clusters of image blocks given by a fixed grid. The annotations of an image are propagated to the image blocks. The blocks of new images are assigned to the clusters and their co-occurrence statistics accumulated for the image. The most likely words are assigned as annotations to the new image. The system in [47] proposes to use machine translation techniques for annotating images and regions with words. In this case, the conditional probability of words given region clusters is obtained through maximum likelihood estimation with the EM algorithm. This work also

proposes ways for refining the lexicon of words by (1) removing words that cannot be predicted with a minimum probability by any region cluster, and (2) clustering similar words using the conditional probability of image regions given the words. In a similar way, the approach presented in [5] consists on modeling the joint distribution of words and region feature clusters of images using Hofmann's hierarchical clustering model [58].

Our image classification framework differs from related work in automatically selecting salient classes and in combining multiple classifiers based on medianets discovered from annotated images that use external knowledge from WordNet. In addition, our work analyses the role of visual and textual features in image classification. The most similar related work to ours is [110] and [104], which build Bayesian Networks (BNs) to classify images or video shots in which some of the nodes in the BN are also classifiers. However, in these approaches, the topology of the BN is either manually entered by experts or automatically learned using costly statistical methods. In contrast, both the target classes and their relationships in our classification framework are automatically obtained through discovering and summarizing medianets from annotated images.

5.3 Building MediaNet Classifiers

The proposed approach for building a MediaNet classifier from a medianet consists of three steps, as shown in Figure 5.1. First, a meta-classifier is trained to detect each concept (of the medianet) in images. Then, a Bayesian Network (BN) is built using the meta-classifiers and the discovered concept network. We present two approaches for

building the Bayesian network: using the meta-classifiers as the only nodes in the network, and having also nodes representing true concepts. In an optional third step, new statistical relationships from the Bayesian network can be learned and added to the medianet. This section describes each step in detail.

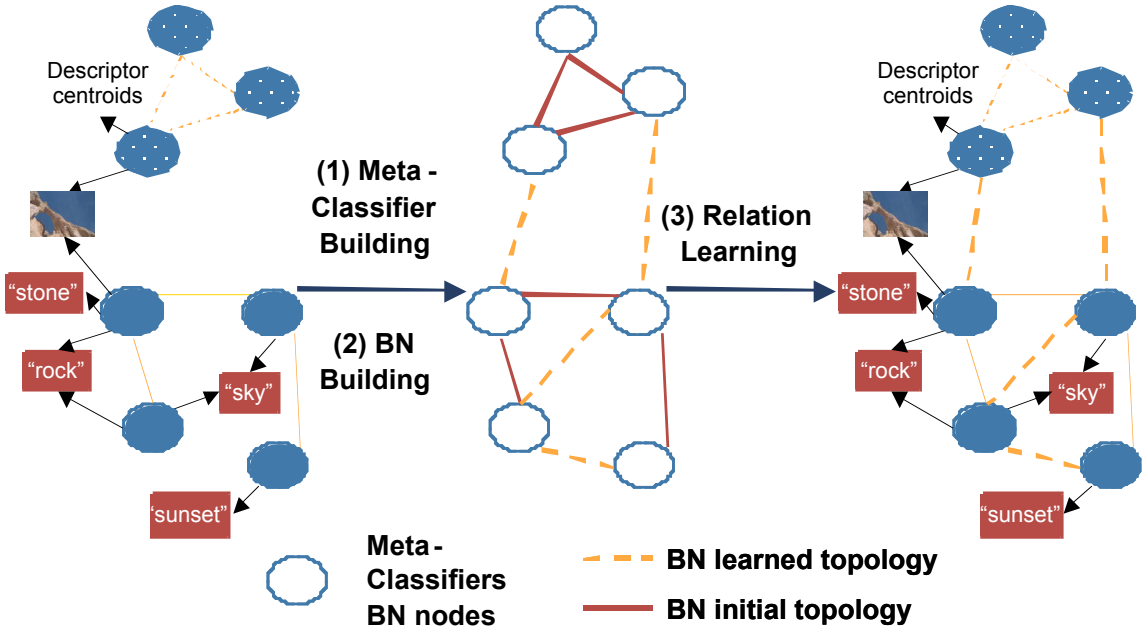


Figure 5.1: The construction of the MediaNet classifier from a medianet consists of three steps: (1) meta-classifiers are trained to predict the present of concepts in images; (2) a Bayesian Network (BN) is built using the meta-classifiers and the concept network; and, optionally, (3) new statistical relationships from the Bayesian network can be learned and added to the initial medianet.

5.3.1 Building Meta-Classifiers

In the first step, meta-classifiers are trained to detect concepts in images based on their visual and textual features. If more than one classifier is trained for each concept in the medianet, the classifiers are combined into a meta-classifier.

A classification algorithm learns how to predict the class (the label of the class attribute) for an input (given feature attributes of the input) [45]. We consider a diverse set of classification algorithms: Naïve Bayes (NB), Support Vector Machine (SVM), and k-Nearest Neighbor (KNN) classifiers. The rationale for selecting each algorithm follows. The Naïve Bayes classifier is one of the simplest classifiers. SVM classifiers are slow at training but quick at classification; in addition, they have shown promising classification performance in various multimedia problems. The KNN classifier can be trained quickly but it is slow at classification. Finally, KNN classifiers do not require large training sets. In the experiments presented in this chapter, we only use NB and SVM classifiers.

A classifier is trained to detect a concept in images. The feature attributes input to each classifier are all or a subset of the visual and textual features (e.g., the Tamura texture and tf * idf feature vectors, respectively) for each image. Feature normalization is done when classifiers deal with multiple features (i.e., the mean and variance of vector bins are normalized to zero and one, respectively). This might not be necessary for some classifiers can do this implicitly. The class attributes are labels such as {presence, no presence} indicating the strength of the presence of a concept in an image. For perceptual concepts or image clusters, the class labels are the presence or absence of an image in the cluster; for semantic concepts or word senses, the quantized word-sense disambiguation scores (see section 4.4.2). Concept occurrences need to be propagated on the medianet for obtaining the class labels of images because an image that shows a dog also shows an animal as a dog is a kind of animal (see section 4.5.1).

In the case of two-class label classifiers (e.g., SVMs), several classifiers are used to learn more than two labels for each concept (e.g., {strong presence, weak presence, no presence}) using the one-per-class coding technique [45]. This technique consists of training one classifier for detecting each label. Multiple classifiers can be trained for the same concept using different combinations of features and classification algorithms (e.g., SVM classifier using Tamura texture and KNN classifier using tf * idf feature vector for a concept). In this case, all the classifiers for a concept are combined into a meta-classifier using techniques such as stacking or majority voting [45]. The NB, SVM, and KNN classification algorithms in Weka [163], a data mining software framework for Java, are used to train and detect concepts in images with or without annotations.

5.3.2 Building the Bayesian Network

The second step consists of building a Bayesian network using the meta-classifiers constructed in the previous step and the medianet. We propose two approaches for building the Bayesian network: using the meta-classifiers as the only nodes in the network, and having also nodes representing each true concept.

Bayesian Networks (BNs), also known as Belief Networks (BNs), are directed graphical models that allow the efficient and compact representation of joint probability distributions of multiple random variables [45]. A Bayesian network is fully specified by the topology of the graph, and the parameters of the conditional probability distribution of each node with respect to its parent nodes. Two reasons prompted the selection of

Bayesian networks for learning statistical dependencies between concepts. First, there are algorithms to learn both the parameters and the topology of a Bayesian network. If the nodes in a Bayesian network represent concepts, then, the algorithms are actually learning statistical relationships among the concepts. Second, once built, the Bayesian network can answer arbitrary probabilistic questions about the concepts (e.g., joint probability for the values of any two nodes), thus functioning as a knowledge classifier in itself.

We propose two approaches for combining meta-classifiers using Bayesian networks (see Figure 5.2). In the first approach, BN:MC (see Figure 5.2.b), the nodes of the BN are the meta-classifiers; each node is thus indirectly representing a concept. The values of the nodes are the predicted class labels of the meta-classifiers, e.g., {strong presence, weak presence, no presence}. The topology of the BN is set to that of the medianet; this is our best guess for the BN's topology based on knowledge discovered using the techniques presented in chapter 4. In the second approach, BN:MC+RC (see Figure 5.2.c), the BN has meta-classifiers and real concepts as nodes; where a real concept node directly represents the true presence of a concept (i.e., independent of the output of the classifiers). The arcs connecting real concept nodes in the BN are the relationships between concepts in the medianet. In addition, real concept nodes have incoming arcs from the meta-classifier nodes associated with corresponding concepts and adjacent concepts in the medianet. We consider the second approach, i.e., BN:MC+RC, because it allows modeling and learning relationships among true concepts as well as between true

concepts and concept labels predicted by classifiers. This provides the potential to improve the overall classification accuracy in spite of the increased network complexity.

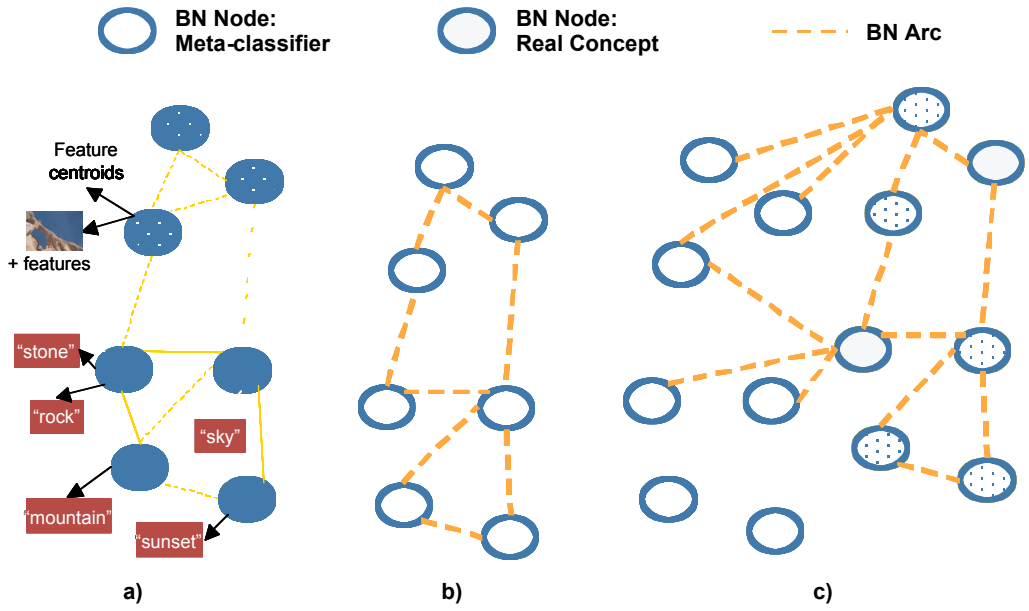


Figure 5.2: Medianet and corresponding MediaNet classifiers: a) medianet; b) MediaNet classifier of meta-classifiers, BN:MC; c) MediaNet classifier of meta-classifiers and real concepts, BN:MC+RC.

As mentioned above, the topology of the BN is derived from the medianet. Each relationship in the medianet is assigned a direction in the BN in accordance with the cause-effect dependencies of a BN, as applicable. For example, a specialization relationship in the medianet translates to an arc from the node corresponding to the specialized concept to the node associated with the general concept (e.g., arc from node "dog" to node "animal" because all the dogs are animals). Bayesian networks cannot have directed cycles so some arcs may need to be removed in the initial topology derived from the medianet. Directed cycles are solved by removing all the arcs between any two

adjacent nodes (i.e., nodes connected by one or more arcs) in each cycle. More advanced strategies could be devised. Once the topology of the BN is set, the parameters of the BN are learned using standard statistical methods. In particular, we do Maximum likelihood parameter estimation using the Bayes Net Toolbox [103].

5.3.3 Learning Statistical Relationships

Optionally, new statistical relationships can be learned in the Bayesian network and added to the existing relations in the medianet. In this case, the parameters are learned together with the structure of the Bayesian network.

We use two approaches for learning the topology of a Bayesian network. The first approach incrementally adds or deletes one arc at a time; whereas the entire topology of the Bayesian network is learned from scratch in the second approach. Although the second approach requires larger training data, in general, learning the topology of a Bayesian network is computationally expensive. We use the Markov Chain Monte Carlo (MCMC) algorithm called Metropolis-Hastings (MH) [56] from the Bayes Net Toolbox [103].

The topology of the learned Bayesian network is compared to the one of the medianet. A Statistical Dependent relationship is added to the medianet for each arc between two nodes in the Bayesian network that does not have a relationship between the corresponding concepts in the medianet. In the BN:MC MediaNet classifiers, we use the meta-classifiers nodes; in the BN:MC+RC MediaNet classifiers, the real concept nodes.

New relationships could be added to the medianet for each arc in the learned Bayesian network; however, some of these statistical dependencies would likely be due to already known relationships between concepts.

5.4 Labeling Images

Once trained, the MediaNet classifier uses the meta-classifiers to detect concepts in new images. This initial prediction is refined using Bayesian inference. In addition, since textual or joint visual-textual features have been shown to outperform visual features in image classification [111], we consider an approach that predicts textual features from visual features in the case of images without annotations. This section describes the proposed image classification and textual feature prediction techniques.

5.4.1 Classifying Images

The first step to classify a new image with a MediaNet classifier is to extract visual (and textual) features from the image (and its annotations), as needed. The features are inputted to the meta-classifiers in the MediaNet classifier for a first prediction of the concept labels (i.e., the presence or absence of each concept). Bayesian inference is then used to refine the initial predictions. The specific procedure differs for BN:MC and BN:MC+RC MediaNet classifiers, as described below.

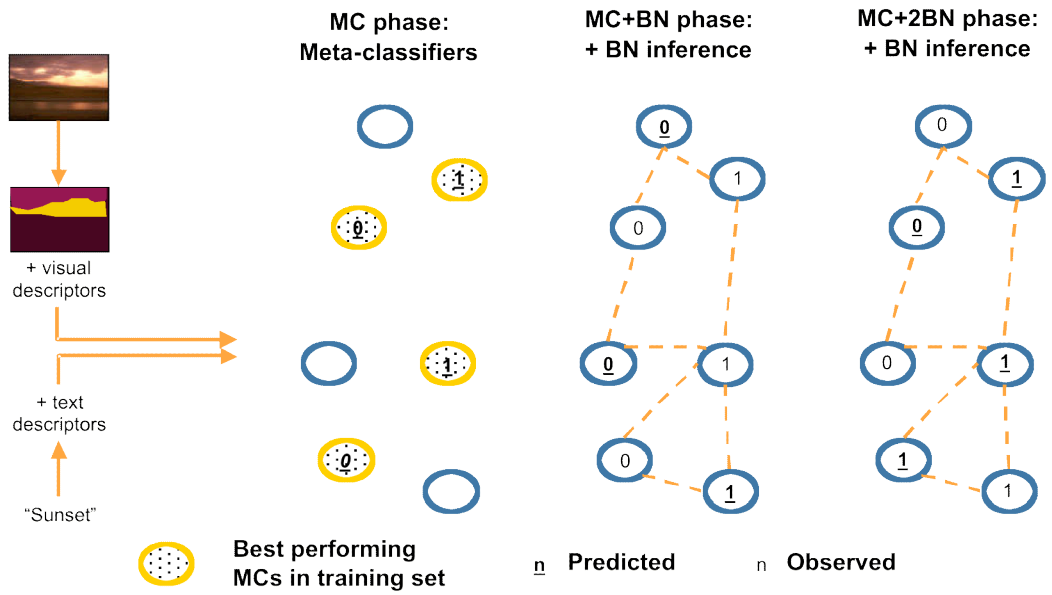


Figure 5.3: Classification procedure for BN:MC MediaNet classifier in Figure 5.2.b: first, the best meta-classifiers are used to detect the concepts of the corresponding BN nodes (MC phase); then, the presence of the other concepts is predicted using Bayesian inference (MC+BN phase); and finally, the presence of the concepts corresponding to the best classifiers can be predicted using Bayesian inference (MC+2BN phase). The circles correspond to concepts in Figure 5.2.a; the number inside a circle is the predicted or observed label of the concept. We consider two possible labels per concept: {1=presence, 0=no presence}.

In a BN:MC MediaNet classifier (see Figure 5.3), the labels of the concepts corresponding to the best meta-classifiers are predicted using the meta-classifiers (MC phase). The predicted concept labels are then entered as observed values of the corresponding nodes in the BN to infer the labels of the remaining concepts using Bayesian inference (MC+BN phase). The best meta-classifiers are found automatically by measuring the performance of each meta-classifier as the concept detection accuracy for the images in the training set. We have tried and evaluated different values for the number of best meta-classifiers to use in the MC phase. Please, note that the labels of

the nodes corresponding to the best meta-classifiers do not change after inference because they are set as observed values in the MC+BN phase. In an optional phase, the labels for concepts detected using the best meta-classifiers can be further refined using the Bayesian network (MC+2BN phase) by considering the labels for the other concepts as observed values of the corresponding nodes in the Bayesian network. Unconnected concepts are labeled using only the meta-classifiers. The final labels of the BN nodes indicate the strength of the presence of each concept in the new image (e.g., either the presence, value = 1, or the absence, value = 0, of each concept in Figure 5.3).

In a BN: MC+RC MediaNet classifier (see Figure 5.4), the output labels of all the meta-classifiers are observed values of the associated meta-classifier nodes in the Bayesian network. Standard Bayesian inference techniques are then used to obtain the labels of the real concept nodes in the network. The presence of each concept is predicted using the inferred label of the corresponding concept node in the Bayesian network. The final labels of the real concept nodes indicate the strength of the presence of each concept in the new image (e.g., either the presence, value = 1, or the absence, value = 0, of each concept in Figure 5.4).

In both kinds of MediaNet classifiers, if the new image has annotations, the senses of the words in the annotations can be disambiguated using the techniques proposed in section 4.4.2. The detected senses in the image annotations can then be set as observed values of the corresponding meta-classifier and real concept nodes in BN:MC and BN:MC+RC MediaNet classifiers, respectively (+O option). The presence of the other

concepts is then predicted following the procedures described above: feature extraction, meta-classifier labeling, and Bayesian inference. The observed values after word-sense disambiguation remain as such during this process.

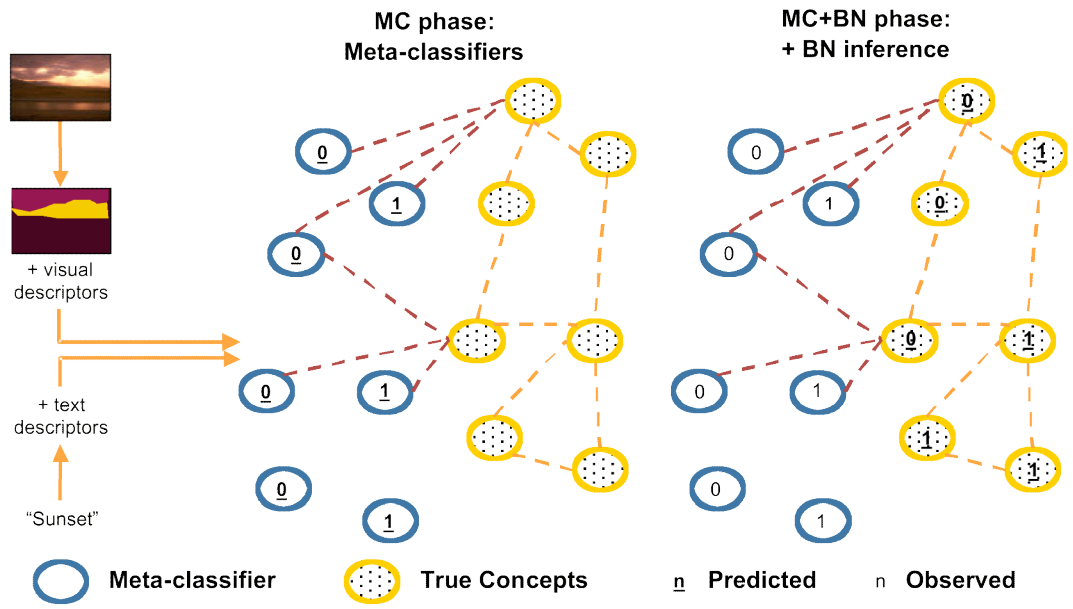


Figure 5.4: Classification procedure for BN:MC+RC MediaNet classifier in Figure 5.2.c: first, the meta-classifiers are used to predict the labels of the corresponding meta-classifier nodes (MC phase); and then, the true presence of the concepts is predicted by the real concept nodes using Bayesian inference (MC+BN phase). The circles filled with dots correspond to concepts in Figure 5.2.a; the number inside these circles is the predicted label of the corresponding concept. We consider two possible labels per concept: {1=presence, 0=no presence}.

5.4.2 Predicting Textual features

If a new image does not have annotations, we predict the textual features from the visual features of the image in order to label the image with MediaNet classifiers that use visual and textual features. We present two different methods for feature prediction that are

based on clustering the textual features and on modeling the visual-textual feature distribution per class, respectively.

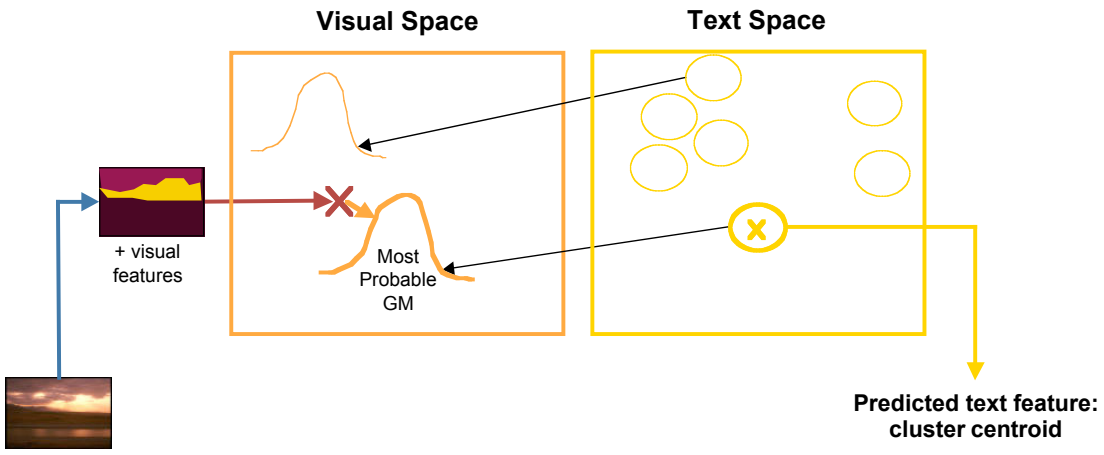


Figure 5.5: Textual feature prediction process based on clustering images using textual features (the circles are clusters) and modeling the visual features of the images within each cluster using a Gausian Model (GM, Gaussian functions). The textual features predicted for a new image with no annotations is the center of the textural feature cluster associated with the most likely Gaussian model given the visual features of the image.

The first proposed method for textual feature prediction consists of clustering the training images based on their textual features using the KNN clustering algorithm (see Figure 5.5). The distribution of the visual features for the images within each cluster is then modeled using a Gaussian model. For a new image without annotations, the most likely Gaussian model is found given its visual features. The predicted textual features for the new image are the centroid of the textual feature cluster for to the most likely Gaussian model. In other words, we use Bayesian detection theory to detect the most likely textual cluster for a new image given its visual features. The propose approach is similar to the mapping between the semantic and the acoustic spaces for audio retrieval

and indexing in [130]. We refer to this approach as the Clustering Prediction approach (CP approach).

For predicting textual features, we also adopt the Statistical Approach proposed (SP approach) for handling unreliable or missing acoustic data presented in [37]. For predicting the textual features given some visual features, the technique in [37] translates into finding the distribution of the textual features x_t conditioned on the visual features x_v and the presence of a concept C, i.e., $f(x_t|x_v,C)$. The distribution of all the features $x = (x_t, x_v)$ in the presence of a given concept C, $f(x|C)$ is assumed to be a mixture of M Gaussian models with diagonal covariances. The textual features given some visual features x_v and the presence of a concept C can be predicted, x'_t, using the expectation of the distribution $f(x_t|x_v,C)$, whose final expression is

$$x'_{t,C} = \sum_{k=1}^{M} P(k|x_v, C)\, \mu_{t,k,C} \tag{5.1}$$

$$P(k|x_v, C) = \frac{P(k|C)\, N(x_v; \mu_{v,k,C}, \sigma^2_{v,k,C})}{\sum_{k=1}^{M} P(k|C)\, N(x_v; \mu_{v,k,C}, \sigma^2_{v,k,C})} \tag{5.2}$$

where $P(k|C)$ and $N(x; \mu_{k,C}, \sigma^2_{k,C})$ denote the coefficient and the Gaussian for mixture component k of $f(x|C)$ in the presence of concept C, respectively; the mean and variance of the mixture component k can be partitioned into textual and visual features as follows: $\mu_{k,C} = (\mu_{v,k,C}, \mu_{t,k,C})$ and $\sigma^2_{k,C} = (\sigma^2_{v,k,C}, \sigma^2_{t,k,C})$, respectively; and

$N(x_v; \mu_{v,k,C}, \sigma^2_{v,k,C})$ corresponds to the Gaussian for the mixture component k for the visual features.

5.5 Evaluation Experiments

In this section, we present the setup and discuss the results of the experiments performed for evaluating the proposed techniques for image classification. We use two data sets: 2706 nature images and 2561 travel images with annotations. A medianet was extracted and summarized for about 80-90% of randomly selected images for each collection. The resulting medianets were used to train and build several MediaNet classifiers with different parameters. The remaining images were used to test and compare the performance of the medianets classifiers in terms of classification accuracy with respect to several baseline approaches such as random classifiers and the meta-classifiers without the Bayesian network. Please, note that the classifiers are built using the summarized medianet instead of the initially discovered medianet. We aim reducing the complexity of building classifiers for the discovered concepts by grouping similar concepts together for.

5.5.1 Setup

We used two different image collections in the evaluation experiments: a nature and a travel image collection. Nature and travel images are common examples of annotated images online and similar to pictures taken by common consumer, especially travel images. This section describes the content and the processing of each image collection.

5.5.1.1 Nature Image Collection

A collection of 2706 nature images was taken from Berkeley's CalPhotos collection (http://elib.cs.berkely. edu/photos). The images in CalPhotos are labeled as plants (857), animals (818), landscapes (660), or people (371). Examples of nature images with annotations are shown in Figure 5.6. We use a few keywords from the annotations describing the main objects or people depicted on the pictures (e.g., "plant, flower"). We separated the 2706 images into two sets randomly: 2437 images for training and 269 images for testing.

Figure 5.6: Examples of nature images. The annotations for one of the images are also provided.

A medianet was constructed using the 2437 training images. Color histogram (166 bins) was extracted from the images; and log tf * entropy (125 bins after latent semantic indexing [42]) from the annotations. Color histogram has been proven to be effective in retrieving natural images [121]; in addition, it is widely accepted that log tf * entropy outperforms other word-weighting schemes for retrieving textual documents [46]. A medianet was then constructed using the correct senses of the nouns in the annotations

and their relationships in WordNet as described in section 4.4. The first author of this thesis generated the ground truth of the correct senses for all the nature images including the training images. The initial medianet, which was purely semantic and composed of 52 concepts, 47 specialization relationships, and 2 aggregation relationships, was summarized into a medianet of 16 concepts and 13 specification relationships as presented in section 4.7.4 using the modified k-NN clustering algorithm. See Table 5.1 for a list of the most frequent words in the annotations, and the concepts in the summarized medianet. Several MediaNet classifiers were then built for the summarized medianet using different classifiers, features, and structures, among others.

Table 5.1: Most frequent words in annotations and concepts in the medianet summary of 16 concepts with occurrence probabilities (in percentages) for the nature images.

Words		**Concepts**	
Plant	15.88	Plant flora, vine, tree	18.66
Animal	15.08	Animal, beast, fauna	14.96
Flower	13.30	Natural object, plant part, flower	13.19
Habitat	12.19	Society, people, group, culture	12.66
Landscape	12.19	Vicinity, country, landscape	12.09
People	6.85	Habitat, geographic area, region	12.09

5.5.1.2 Travel Image Collection

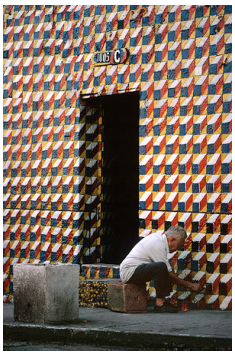
We also employ a collection of 2561 travel photographs. The images were taken by three amateur or almost professional photographers in their travels around the world: Angus

McIntyre [90], Bill Hocker [57], and Martin Wierzbicki [161]. Examples of travel images with annotations are shown in Figure 5.7. We picked and collected these pictures because of our interest in understanding and modeling visual content from consumers [108]. The pictures are similar to the ones taken by common tourists during their vacation travels. In addition, these pictures are annotated with short descriptions from the photographers. Finally, the photographers travel to many of the same countries and continents. We separated the 2561 images into two sets randomly: 2048 images for training and 513 images for testing.

Two girls dressed in white costumes.
Carnevale di Venezia, Venice
1990

Fine rills on the face of a sand dune
Erg Chebbi near Merzouga, Morocco
2001

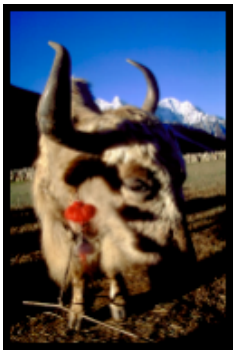
a)



House Painter
Mexico
1973

Harvestor
Korea
1977

b)



Yak at Dingboche (4410m)
Everest Region, Nepal
May 2001

Sunset in Puerto Natales
Puerto Natales, Chile
December 2000

c)

Figure 5.7: Examples of travel images with annotations taken by a) Angus McIntyre [90], b) Bill Hocker [57], and b) Martin Wierzbicki [161].

A medianet was constructed using the 2048 training images. Color histogram (166 bins) was extracted from the images; and log tf * entropy (125 bins after latent semantic indexing [42]) from the annotations. The senses of the words in the annotations were

automatically disambiguated as described in section 4.4.2 using clusters based on color histogram. A medianet was then constructed using the disambiguated senses of nouns in the annotations and their relationships in WordNet as described in section 4.4. The initial medianet, which was purely semantic and composed of 3037 concepts and 3061 specialization relationships, was summarized into a medianet of 16 concepts as presented in section 4.7.4 using the spectral clustering algorithm (although the resulting medianet had only 14 concepts). See Table 5.2 for a list of the most frequent words in the annotations, and the concepts in the summarized medianet. Several MediaNet classifiers were then built for the summarized medianet using different classifiers, features, and structures, among others.

Table 5.2: Most frequent words in annotations and concepts in the medianet summary of 16 concepts with occurrence probabilities (in percentages) for the travel images.

Words		**Concepts**	
Tunisia	2.06	Object, cause, primate	96.90
California	2.03	Something, plant, animal	96.55
Kong	1.38	Location, courtyard, district	88.81
China	1.37	Capital, surface, river basin	88.57
Hong	1.34	Artifact, machine, representation	48.25
Italy	1.12	Article, tower, pot, sanctuary	41.40
View	1.04	Person, god, child, rider, plant	35.34
Japan	0.97	Social relation, communication, time	29.03
Thailand	0.73	Psychological feature, blockade	27.85

5.5.2 Results

We tested and compared the performance of MediaNet classifiers for both image collections with respect to several baseline approaches using the mean classification accuracy for the 16 (super) concepts in the summarized medianet. For a concept, we measured the classification accuracy for each concept as the percentage of images in the testing set to which the concept was correctly assigned. As a concept in the MediaNet classifier can include several elementary concepts (senses, in this case), we consider that a concept is present in an image if, at least, one elementary concept is correctly detected in the image. The classification accuracies per concept were weighted by 1 – p log (p), where p was the probability of the concept in the training annotations. Very common and very rare concepts are, therefore, given lower weights because they are less important. We manually generated the ground truth of correct senses for words in the annotations of the images in the testing sets. The correct senses (elementary concepts) were used to determine which concepts in the medianet summary were present in the testing images.

Table 5.3 and Table 5.4 lists the mean classification accuracy for the nature and the travel images, respectively, using MediaNet classifiers built for (1) different features: color histogram (CH), log tf * entropy (LE), predicted log tf * entropy using the clustering (CPLE) and statistical (SPLE) approaches, and combinations of these (see section 5.4.2); (2) different meta-classifiers (in this case, individual classifiers): SVM and NB; (3) different structures for the Bayesian network: the meta-classifiers with no BN (MC no

BN), BN of meta-classifiers (BN:MC), and BN of meta-classifiers and real concepts (BN:MC+RC) (see section 5.3.2); (4) and learning the parameters (PA) and also the structure (+ST) of the BN. The accuracies for BN:MC correspond to the best MediaNet classifier at the MC+BN or the MC+2BN phase using 1, 2, 6, or 8 of the best performing meta-classifiers in the MC phase (see section 5.4.1). In addition, we include results from observing values of nodes in the BN based on disambiguated senses in the annotations of the testing images (+O) (see section 5.4.1). For baseline comparison, randomly deciding the presence of concepts in images resulted in accuracies of about 50%.

Table 5.3: Mean classification accuracy for the nature images using different classifiers (SVM: Support Vector Machines, NB: Naïve Bayes), different input feature features (CH: color histogram, LE: log tf * entropy, CPLE: LE predicted using clustering approach, SPLE: LE predicted using statistical approach), different structures of the BN (MC no BN: only meta-classifiers without the BN, BN:MC: BN of meta-classifiers, BN:MC+RC: BN of meta-classifiers and real concepts). Columns PA and + ST are results for learning the parameters, and also the structure of the BN, respectively. Column +O are results from observing nodes in the BN+PA case for senses disambiguated in annotations.

	CH						
	MC	**BN:MC**			**BN:MC+RC**		
	no BN	**PA**	**+ST**	**+O**	**PA**	**+ST**	**+O**
SVM	84.31	81.93	83.00	84.36	82.30	80.40	94.27
NB	65.45	65.3	60.56	68.89	81.33	80.40	94.96

	CH+CPLE			CH+SPLE		
	MC no BN	**BN: MC**	**BN: MC+RC**	**MC no BN**	**BN: MC**	**BN: MC+RC**
SVM	79.30	79.19	79.40	43.40	66.32	39.90
NB	70.35	77.24	78.94	60.79	75.68	77.16

	LE						
	MC no BN	**BN:MC**			**BN:MC+RC**		
		PA	**+ST**	**+O**	**PA**	**+ST**	**+O**
SVM	99.66	95.72	99.66	99.56	99.74	83.12	99.81
NB	85.52	83.87	83.88	87.74	90.35	83.31	95.48

	CH+LE						
	MC no BN	**BN:MC**			**BN:MC+RC**		
		PA	**+ST**	**+O**	**PA**	**+ST**	**+O**
SVM	99.58	95.59	95.58	99.48	99.61	83.09	99.62
NB	82.76	82.10	81.10	88.29	86.04	80.47	92.40

Table 5.4: Mean classification accuracy for the travel images using different classifiers (SVM: Support Vector Machines, NB: Naïve Bayes), different input feature features (CH: color histogram, LE: log tf * entropy, CPLE: LE predicted using clustering approach, SPLE: LE predicted using statistical approach), different structures of the BN (MC no BN: only meta-classifiers without BN, BN:MC: BN of meta-classifiers, BN:MC+RC: BN of meta-classifiers and real concepts). Columns PA and + ST are results for learning the parameters, and also the structure of the BN, respectively. Column +O are results from observing nodes in the BN+PA case for senses disambiguated in annotations.

	CH						
	MC no BN	BN:MC			BN:MC+RC		
		PA	+ST	+O	PA	+ST	+O
SVM	81.67	81.80	81.80	89.68	81.65	81.65	91.23
NB	50.67	50.11	48.70	62.80	80.78	80.80	90.74

	CH+CPLE			CH+SPLE		
	MC no BN	BN: MC	BN: MC+RC	MC no BN	BN: MC	BN: MC+RC
SVM	81.10	81.80	80.86	71.86	81.78	71.84
NB	63.46	63.44	80.86	50.06	60.04	81.55

	LE						
	MC no BN	BN:MC			BN:MC+RC		
		PA	+ST	+O	PA	+ST	+O
SVM	88.11	82.58	82.58	90.34	88.15	88.10	93.77
NB	76.21	79.67	79.67	87.49	84.29	84.32	91.68

	CH+LE						
	MC no BN	BN:MC			BN:MC+RC		
		PA	+ST	+O	PA	+ST	+O
SVM	88.40	82.52	82.50	90.41	88.41	88.39	93.68
NB	62.25	60.64	60.64	60.64	81.61	81.71	90.97

5.5.3 Discussion

The classifiers for the nature images (see Table 5.3) use the correct senses of words in annotations during the medianet and classifier construction. We do this for the purpose

of decoupling classification and disambiguation errors. If senses were disambiguated automatically for the nature images, as described in section 4.4.2, only 65% of the words would have been disambiguated correctly. However, classification accuracies still reached 90% and 80% for SVM and NB, respectively, using log tf * entropy features and color histogram + log tf * entropy features. In addition, for both, correct and automatically disambiguated senses, we observed similar trends in the results for the same features, classifiers, etc. For the travel images (see Table 5.4), we report on classifiers that use automatically disambiguated senses. The disambiguation accuracy for the travel images in the testing set was above 84% for nouns.

For the nature images, as shown in Table 5.3, if annotations are available for new images, the best performing systems use (1) the individual SVM meta-classifiers (MC no BN) and (2) the BN of SVM meta-classifiers and real concepts (BN: MC+RC), using either textual features (LE) or textual-visual features (CH+LE). However, the difference in accuracy of these systems is not significant. When annotations are not available for classification (i.e., only color histogram inputs to meta-classifiers), the highest accuracy is achieved again for (1) the individual SVM meta-classifiers and (2) the BN of SVM meta-classifiers and real concepts.

For the travel images, as shown in Table 5.4, if annotations are available for new images, the best performing systems use (1) the BN of SVM meta-classifiers and real concepts with observed values (SVM, BN: MC+RC, +O) and (2) the BN of NB meta-classifiers and real concepts with observed values (NB, BN:MC+RC, +O), using either textual

features (LE) or textual-visual features (CH+LE). The accuracy of these systems is about the same. When annotations are not available for classification (i.e., only color histogram features are inputted to meta-classifiers), the highest accuracy is achieved for (1) the individual SVM meta-classifiers and (2) the BN of SVM meta-classifiers with or without real concepts. Again, the performance of these systems is comparable.

For both image collections, and using annotations or not, having real concepts in the BN (BN:MC+RC) generally outperformed the BN of meta-classifiers alone (BN:MC) by up to 15%. This is a good indication of the importance of including nodes corresponding to real concepts in the BN. Although the improvements for the BN of meta-classifiers and real concepts are not very significant with respect to single SVM classifiers (up to 5%), gains of up to 15-30% in accuracy where obtained for NB classifiers. For the travel images, the improvement of the MediaNet classifiers over single classifiers was more widespread for different features and parameters: up to 4% with SVM classifiers and up to 20% for NB classifiers. The effect of the MediaNet classifiers was to balance the performance of classification frameworks built with different features and individual classifiers. Therefore, combining classifiers using a BN can offer significant performance gains that are not affected by specific choices of features and classifiers. In the case of classifiers that originally provide good classification accuracies, the gains may not be very significant (4% for SVM classifiers for travel images).

Other conclusions can be drawn from Table 5.3 and Table 5.4. First, the structure of the medianet discovered from annotated images using WordNet helps in labeling new

images. BNs of meta-classifiers (especially with real concepts) whose structures were based on discovered medianets consistently outperformed BNs with purely statistically learned structures by up to 15% for the nature images for both SVM and NB classifiers. In addition, observing values of nodes in the BN based on disambiguated senses in annotations (+O) improves the accuracy and robustness of MediaNet classifiers even with textual feature inputs. As an example, the most accurate NB-based MediaNet classifier for nature images used color histogram inputs and observed values of nodes in the BN based on disambiguated senses. For travel images, the best performing SVM-based MediaNet classifier also observes values of BN nodes based on disambiguated senses. Finally, predicting textual features using visual features did not improve the most accurate MediaNet classifier with color histogram inputs with the SVM classifier, although their performance is comparable. However, it improved the results of meta-classifiers without the BN and of the BN of meta-classifiers for the NB classifier.

5.6 Summary

This chapter has presented novel methods for classifying images based on knowledge discovered from annotated images in the form of medianets. The novelty of this work is the automatic class discovery and the classifier combination using extracted medianets.

The extracted medianets are networks of concepts (e.g., image clusters and word-senses) with associated text and images, which are built automatically from annotated images using the electronic dictionary WordNet. Concepts that are similar statistically are

merged to reduce the size of the medianet. Our MediaNet classifier is constructed by training a meta-classifier to predict the presence of each concept in images. A Bayesian network is then learned using the meta-classifiers and the concept network. For a new image, the presence of concepts is first detected using the meta-classifiers and refined using Bayesian inference.

Experiments have shown that classifiers based on medianets discovered and summarized from annotated images using human knowledge (i.e., WordNet) can result in superior accuracy to individual classifiers and purely statistically learned classifier structures. The improvement in accuracy for classifiers that are not originally very good is significant (up to 15-30% for NB classifiers); however, the gains for originally accurate classifiers are more modest (up to 4% for SVM classifiers). Another contribution of this work is the analysis of the role of visual and textual features in image classification. As textual or joint visual-textual features perform better in classifying images than visual features, we try to predict textual features for images without annotations; however, we have found that the accuracy of visual-predicted textual features does not consistently improve over using only visual features.

Many times the labels or the words associated with images only apply to specific regions of the images. Imagine the picture of a tiger in the wild; the label "tiger" is only applicable to the region that depicts the tiger. In the future, we plan to develop algorithms for distinguishing between concepts that are applicable to entire images (e.g., "outdoor") or only to regions within images (e.g., "tiger"). We are also interested in the

problem of classifying image regions using annotations assigned globally to images. Some initial work in this area can be found in the literature [5][47][106].

6 Knowledge-Based Image Browsing

6.1 Introduction

This chapter focuses on novel methods for browsing images based on medianets constructed from annotated images using WordNet. Current approaches organize images in complex or low level structures that result in inefficient and counterintuitive navigation. In chapter 4, we presented our techniques for automatically discovering and summarizing medianets from annotated images in the form of image clusters, words senses, and relationships among them.

In recent years, there has been a major increase in available multimedia and in technologies to access multimedia. Users need and want tools for effectively and efficiently organizing and browsing multimedia, preferably, at the semantic level (e.g., people and objects in multimedia). Through browsing, users can gain a quick insight into the content of a collection and perform a variety of exploration tasks, with or without a particular goal in mind (e.g., finding a specific image, finding relevant images to a specific topic, or answering some questions). According to Lin [84], browsing is superior to searching when (1) there is a good underlying structure among items so that the similarity between items can be inferred; (2) users are unfamiliar with the collection and

the organization structure; and (3) it is difficult for users to verbalize their information need; among others.

This chapter focuses on the organization and browsing of collections of images with annotations. Current multimedia browsing approaches are often based on low-level visual features (e.g., color and texture) failing to meet user needs at semantic levels. Moreover, they organize images in a single structure, space, or network that is usually too restricted or too complex for non-expert users to navigate. In addition, many current systems lack adequate advanced interfaces for visualizing multimedia. Finally, to date, there have not been systematic evaluations that rigorously assess the effectiveness and efficiency of these browsing systems through extensive user-centered, task-oriented studies.

In this chapter, we present innovative approaches for multi-resolution organization and browsing of images using both visual and textual features [13]. The main contribution of this work is the organization of annotated images in hierarchies of medianets based on medianets extracted from both the images and the annotations, as described in chapter 4 (see Figure 6.1). In addition, we propose to browse the image collection by navigating the medianet hierarchy in a non-strictly hierarchical fashion and by using advanced visualization techniques such as fish-eye views. Finally, we have conducted a user study that extensively evaluates the proposed techniques by measuring the effectiveness, the efficiency, and the subjective satisfaction of users in carrying out common browsing

tasks such as searching for images related to a specific topic. We use the term "MediaNet browser" to refer to our image browsing system.

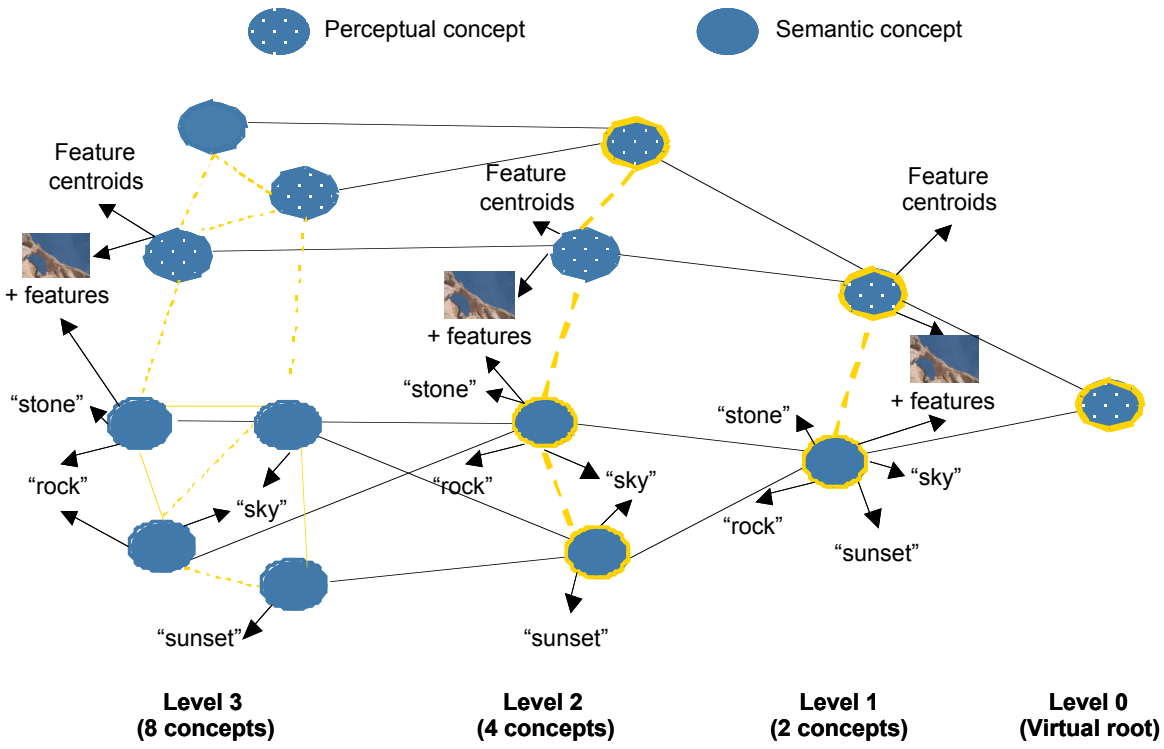


Figure 6.1: Multi-resolution hierarchy of three medianets (with two, four, and eight concepts, respectively) and the root. A concept at a resolution level has two children at the higher resolution level except for leaf nodes, which have no children.

We organize images using a medianet extracted from the collection as described in chapter 4. Then, we iteratively merge statistically similar concepts to summarize the initial medianet at different resolutions. This process results in medianets at different levels of detail forming a hierarchy, as shown in Figure 6.1. For example, the hierarchy in this figure has three medianets (with two, four, and eight concepts, respectively) and the root. Users can browse the image collection by navigating the medianet hierarchy. The

visualization interface follows the Visual Information-Seeking Mantra [128]: "Overview first, zoom and filter, then details-on-demand". The typical zoom in/out and pan next/previous concept operations are supported together with more advanced browsing operations that display concepts, images, and their relationships, among others. The medianet hierarchy can be navigated in a non-strictly hierarchical fashion using our MediaNet browser (e.g., display similar concepts to child concepts).

In addition, we propose novel and advanced techniques for visualizing medianet hierarchies. We use fish-eye views (similar to [38], [50], [51], [148], etc.) for displaying concepts in medianets using representative text and images (e.g., change size of images while displaying concepts based on the concept frequency). A medianet (or part of a medianet) is drawn on a 2D display using spring modeling [75] giving users additional information about concept similarity (e.g., more similar concepts are closer on the display). At any given point, a network with at most eight concepts is displayed to the user because that is the about the size of the human short-term memory [96]. Finally, the concept hierarchy is displayed as a browsable index tree of text and images.

We have conducted an extensive task-oriented, user-centered study to evaluate the proposed techniques in terms of subjective satisfaction, efficiency, and effectiveness of users in performing common browsing tasks. The main task given to subjects was to find relevant images for travel pamphlets related to specific topics. We have compared the performance of our approach with the thesaurus-based image browsing system proposed by Yang et al. [165] through the administration of questionnaires to subjects

and the monitoring of system parameters in the background (e.g. execution time). In particular, we found the MediaNet browser was significantly more useful, easier to use, more stimulating, and more successful. We also found that users needed to see fewer images and to execute fewer browsing operations to performed the simulated tasks.

6.1.1 Outline of the Chapter

The rest of the chapter is organized as follows. In section 6.2, we review some related work on image organization and browsing, and on the evaluation of these techniques. In section 6.3, we summarize the multi-resolution organization of annotated images in hierarchies of medianets. In section 6.4, we describe the proposed MediaNet browser for browsing image collections. We present the proposed visualization techniques for image browsing by navigating medianet hierarchies in section 6.5. In section 6.6, we present the user study conducted for evaluating the proposed image browsing techniques. Finally, we conclude with a summary of the chapter and a discussion of future work in section 6.7.

6.2 Related Work

The problem of browsing images has not received as much attention as retrieving images in the literature. In this section, we review prior work on image browsing in terms of image organization, visualization techniques, and system evaluation. Some methods do not use any structure to organize images; they simply map images onto one- or two-dimensional spaces based on visual and/or textual features [38][120]. However, most methods organize images in specific structures to enable efficient organization of large

image collections. In this case, images are usually organized in clusters based on visual and/or textual features [5][31], or in concept networks based on existing thesaurus [144][165]. Regarding the visualization of images, many systems just show images in low dimensional spaces [38][120], or the ones associated with clusters [31] or concepts [165] in a hierarchical structure. More recent work proposes the use of more sophisticated visualization techniques such as fish-eye views (similar to [38], [50], [51], [148], etc.) and visual display of networks [144] for providing additional information about image similarities (e.g., the closer the images, the more similar). Advanced visualization techniques are more widely used in browsing textual documents [50]. Although, the main goal of browsing systems is to assist users in exploring unfamiliar collections; most proposed methods for image browsing have not been evaluated with actual users. Recently a trend has started for evaluating not only browsing but also retrieval systems using task-oriented and user-centered studies [72][88][153]. In this section, we discuss these and other related work in detail.

Early work on image browsing organizes and displays images into low dimensional spaces based on visual features. For example, MacCuish et al. [89] use Multidimensional Scaling (MDS) to browse the images returned by queries in a two-dimensional plane. Rubner et al. [120] apply MDS to organize all the images in a collection. Santini and Jain [125] also display the images returned by queries in a 2D plane; however, users can iteratively manipulate the position of the images on screen to better characterize the retrieval goal. Similarly, Pecenovic et al. [113] also integrate browsing and retrieval by

displaying the results of queries and all the images in a 2D plane that users can navigate using zooming and panning operations. Images are projected in the 2D plane using Principal Component Analysis (PCA) and Sammon's Projection (SP) based on visual features extracted from the images. These methods run into display problems because images can overlap on the screen. Tian et al. [147] use PCA to visualize images on a 2D plane based on visual and textual features; however, they also propose an optimization strategy to adjust the position and size of the images on screen to minimize overlap while maintaining fidelity of mutual similarities. All these methods do not impose any specific structure in organizing images so they can only effectively handle a small number of images.

More recent methods address this issue by organizing the images into hierarchical clusters that provide visualizations of the images at different levels of detail. Zhang and Zhong [170] cluster a collection of images using the Self-Organizing Map (SOM) algorithm based on visual features and aggregate the clusters to form a hierarchy. A representative image is selected and displayed for each cluster for browsing purposes. Chen et al. [31][32] also organize images in hierarchical clusters; however, they propose an efficient implementation of the standard agglomerative clustering algorithm that reduces memory and computation requirements. In addition, the same authors propose to extend the system to enable users to modify dynamically the organization of the images according to a selected set of relevant images [33]. Similarly, the STELLA System [62] provides an interactive environment that enables users to modify the hierarchical

clusters of images to build digital albums. In the STELLA System, clusters can be removed, merged, or split, among others. Other systems that organize images in hierarchical clusters that users can modify are PICSOM [81] and PIBE [7]. More recently, Stan et al. [140] propose to organize images in hierarchical clusters using MDS techniques but this approach may also suffer from the image overlap problem described above. Vendrig et al. [156] do not explicitly cluster images; however, their strategy to browse images works in a similar fashion. Users select images from different screens until convergence to a small set of images. The main problem with these systems is that the hierarchical clusters are generated based on low-level visual features so the image organization reflects low-level similarity among images. However, users usually want to retrieve, browse, and filter images using semantics instead [1][73]. User interaction only addresses this problem partially because personalized image hierarchies for specific users cannot usually be reused for other users.

Prior work that supports more semantic-like image browsing organizes images using semantic categories generated manually or semi-automatically by experts. Apart from retrieving images based on visual features, WebSEEk [136] supports subject hierarchy browsing of images crawled from the web with related text (e.g., labels, page title, and words in url). The subject hierarchy is built using directory names and frequent words in a semi-automatic process. Mojsilovic and Gomes [99] build visual features for detecting semantic classes in medical images. The semantic classes are decided and organized hierarchically by experts. Related work organizes images using categories from existing

classification schemas or thesauri. Tansley [144] uses a (manually selected) subset of the Dewey Decimal Classification (DDC) schema for annotating and organizing museum images in a concept network. Yang et al. [165] use a sub-hierarchy of the electronic dictionary WordNet generated from the nouns in image annotations entered by users. For each noun, the most common sense of the word is added to the subject hierarchy (if it does not exist already). In addition, the most specific common ancestor in WordNet of the new sense and any sense in the subject hierarchy is inserted in the hierarchy (if it does not exist already; see Figure 6.9.a and Figure 6.9.b). The main drawback of these approaches is that the subject hierarchy (or concept network) organizing the images may become very large and complex. Non-expert users will likely find such large subject hierarchies or concept networks confusing, counterintuitive, and ineffective for browsing, as we found in our study. There are interactive and flexible tools that try to address these issues. For example, the tool [76] visualizes the concept hierarchy in WordNet from the vantage point of a particular word in the dictionary with a given radius of related words. However, the visualization is limited to text; no images are used to help the exploration.

In addition, the visualization of the images (and, if applicable, the image organization structures such as subject hierarchies) in the techniques discussed above is usually rudimentary. Images are displayed in low-dimensional spaces as icons, which may overlap, as mentioned above. In the most advanced systems, users can pan and zoom, or have overview diagrams while browsing the images [113]. Approaches that organize images in hierarchical clusters usually visualize multiple clusters in a 2D grid and display

each cluster with the centroid(s) of the cluster. Users may see all the clusters at one level of the hierarchy, or may zoom in or out of selected clusters, depending on the system. Finally, thesaurus-based approaches usually support hierarchical browsing of the subject categories. The images are shown for the leaf nodes or for any node in the hierarchy, depending on the system. Tansley [144] uses a simple and predefined graphical visualization of a few connected concepts. More sophisticated visualization techniques make use of fish-eye views (e.g., change size and other attributes of images based on current viewing point or interest) to represent similarity or distance relationships among the images (similar to [38], [50], [51], [148], etc.). There is also work that proposes more advanced visual structures for image browsing such as Pathfinder networks [34] and spiral/concentric rings [148]. These advanced visualization techniques are more widely used in browsers of textual documents [50][169]. The Information Navigator [50], for example, displays textual documents and keywords graphically as networks built from occurrence statistics. It also supports fish-eye views and overview diagrams to facilitate the navigation of the networks. Similarly, Zhang and Mostafa [169] extract key terms from textual documents using a vocabulary generation algorithm. The key terms are displayed in a network layout with different colors to indicate term frequency. The system can also present clusters of terms to reduce display clutter.

The main goal of browsing systems is to assist users in exploring unfamiliar collections; however, most methods mentioned above have not been evaluated with real users. Exceptions are [113] and [165], which report results from some user studies. Recently the

trend is to evaluate not only image browsing systems but also image retrieval systems using task-oriented and user-centered studies. Combs and Bederson [36] study the advantages of zoom operations to improve image browsing. Rodden et al. [118] investigate whether it benefits users to arrange thumbnails according to their visually similarity, so images that are alike are placed close to each other. Markkula et al. [88] and Urban et al. [153] propose a test collection and compare several approaches for evaluating content-based image retrieval systems in task-oriented and user-centered studies. The evaluation of browsing techniques is challenging due to the wide range of possible tasks (e.g., locate, distinguish, and categorize images) in addition to the complications of interacting with users. To help in this respect, Shneiderman [128], Morse et al. [101], and Yang-Pelaez [166] propose the use of a task by data type taxonomy, a user task taxonomy, and several information theory metrics for designing and evaluating information visualizations and displays, respectively.

This chapter presents innovative image browsing techniques, which differ from related work as follows. Annotated images are organized in medianets or concepts networks automatically built from the images and the annotations. The medianets include knowledge from the electronic dictionary WordNet. The main difference with the thesaurus-based approaches mentioned above is that the extracted concept network is automatically summarized at multiple levels of detail to facilitate efficient browsing. In addition, we propose to incorporate advanced navigation and visualization techniques, similar to the ones used with textual documents. These include fish-eye views, graphical

visualizations, and non-strictly hierarchical navigation. Furthermore, we evaluate and compare the subjective satisfaction, the effectiveness, and the efficiency of the proposed techniques with respect to prior work in an extensive user study. In the study, users were asked to perform and evaluate different aspects of common browsing tasks such as searching for images related to specific topics. Several system parameters were also recorded and monitored during the user study. We compared the MediaNet browser with a baseline system using a conventional thesaurus-based hierarchical structure [165].

6.3 Multi-Resolution Image Organization

In this chapter, we propose to organize annotated images using hierarchies of medianets. Medianet hierarchies are constructed from knowledge extracted and summarized for the annotated images using the electronic dictionary WordNet (see chapter 4).

For a collection of annotated images, we build the medianet at the highest resolution through the discovery of image clusters, word senses, and relationships between them as described in chapter 4. We construct the medianets at lower resolutions by iteratively clustering statistically similar concepts together (knowledge summarization). This process results in a hierarchy of medianets at multiple resolutions or levels of detail, as shown in Figure 6.1. The perceptual knowledge discovery, the semantic knowledge discovery, and the knowledge summarization techniques were described in chapter 4.

We build the medianet hierarchy so that the medianet at the lowest resolution level has n concepts, and each concept at a resolution level groups at most m concepts of the higher

resolution level. We add an additional virtual root to the medianet hierarchy that includes all n concepts at the lowest resolution level (see Figure 6.1). We cluster the concepts in a medianet using the spectral clustering algorithm [107] as opposed to the modified KNN clustering algorithm (see section 4.5.2). The spectral clustering algorithm was selected because the resulting concept hierarchies were considerably more balanced (i.e., each concept cluster groups a similar number of concepts) than concept hierarchies built using the modified KNN clustering algorithm and, therefore, much shorter. The most frequent concept(s) of a concept cluster is considered to be the centroid(s) of the cluster.

In this chapter, the probability of a concept is calculated as a linear combination of the concept probabilities before and after propagation of the concept occurrences (see section 4.5.1) as follows:

$$p'(c) = a * p(c) + (1-a)\, p_o(c) \tag{6.1}$$

where p(c) is obtained using equation (4.5) and $p_o(c)$ is calculated as follows:

$$p_o(c) = \frac{freq_o(c)}{\sum_{c \in concepts(N)} freq_o(c)} \tag{6.2}$$

where $freq_o(c)$ is the strict concept frequency of concept c before propagation, as defined in 4.5.1. We use concept probability p'(c) as a measure of the importance of

concepts: the more probable the concept, the more important. Our approach takes into account both the strict frequency of concepts in annotations and the frequency after propagation in the concept network. In addition, we can assign different weights to each of these frequencies.

6.4 Browsing Images Using Medianets

Users can browse a collection of annotated images by navigating the medianet hierarchy constructed from the annotated images using the proposed MediaNet browser. The proposed techniques for navigating medianet hierarchies follow the Visual Information-Seeking Mantra [128]: "Overview first, zoom and filter, then details-on-demand". The typical zoom in/out and pan next/previous concept operations are supported together with more advanced browsing operations that display the concepts associated with any image, among others. The medianet hierarchy can be navigated the structure in a non-strictly hierarchical way using the MediaNet browser. This section describes the user interface of the MediaNet browser and the proposed techniques for medianet hierarchy navigation.

6.4.1 The MediaNet Browser

The interface of the MediaNet browser is shown in Figure 6.2. The interface follows the layout of popular hierarchical data browsers such as Windows Explorer for directories and files, and Yang et al. [165]'s browser for words and images.

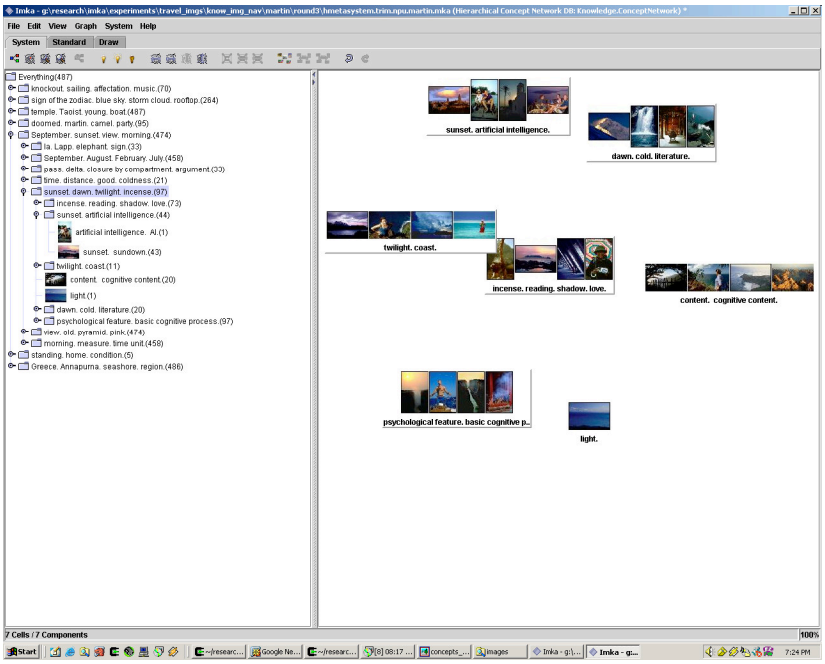


Figure 6.2: Screen shot of the MediaNet browser illustrating the navigation of medianet hierarchies and the system interface. The screen is divided in two windows: a global view of the concept hierarchy is provided on the left side; whereas, a local view of the selected concept can be seen on the right side.

The MediaNet browser provides two views of the medianet hierarchy side to side. The first view is a global view of the medianet hierarchy in the form of a concept hierarchy or tree (see left side of Figure 6.2). The second view is a local view corresponding to the visualization of the selected concept, of the images belonging to a concept, or of the concepts to which an image belongs (see right side of Figure 6.2, Figure 6.3, or Figure 6.4, respectively). Most of the navigation operations discussed in section 6.4.2 can be executed on any of the two views of the medianet hierarchy. Both views are synchronized in such a way that the concept hierarchy shows and highlights the selected

concept at any time, if any. In other words, if the user changes the selected concept in the local view, the global view is updated, and vice versa.

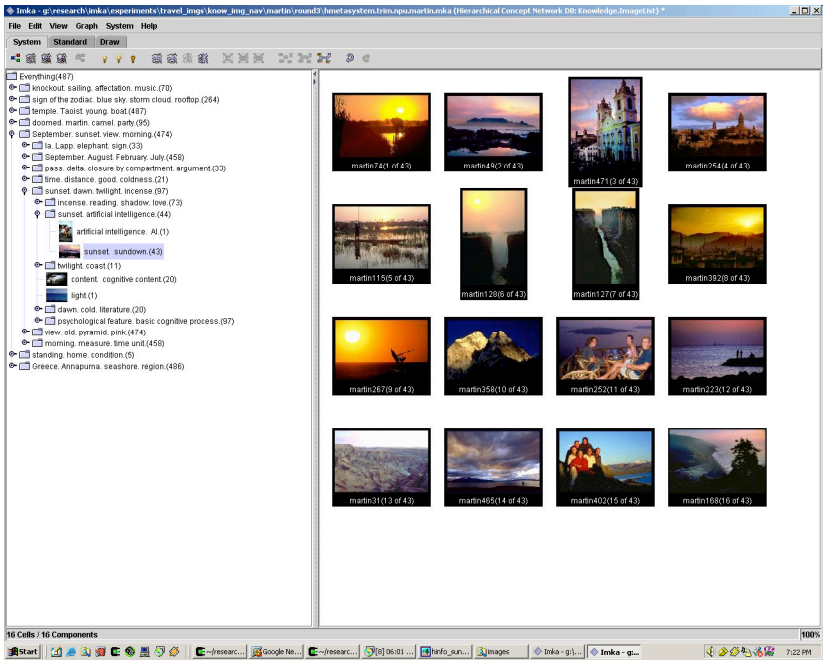


Figure 6.3: Screen shot of the MediaNet browser illustrating the visualization of the images associated with a concept on the local view window (right side).

The window corresponding to the local view visualizes data such as images and concepts as graphs. Users can drag images and concepts around in the local view window. In addition, facilities are provided for deleting, pasting, and copying these nodes from or to other windows in a similar way to the workings of Word Processor. The interface is, therefore, flexible enough to capture sophisticated feedback from users to modify, personalize, and optimize the medianet hierarchies in a much more powerful way than related systems that only allow users to modify clusters [33][62][7][81] or layouts [125] of

images. However, the feedback mechanism is not currently implemented in the MediaNet browser.

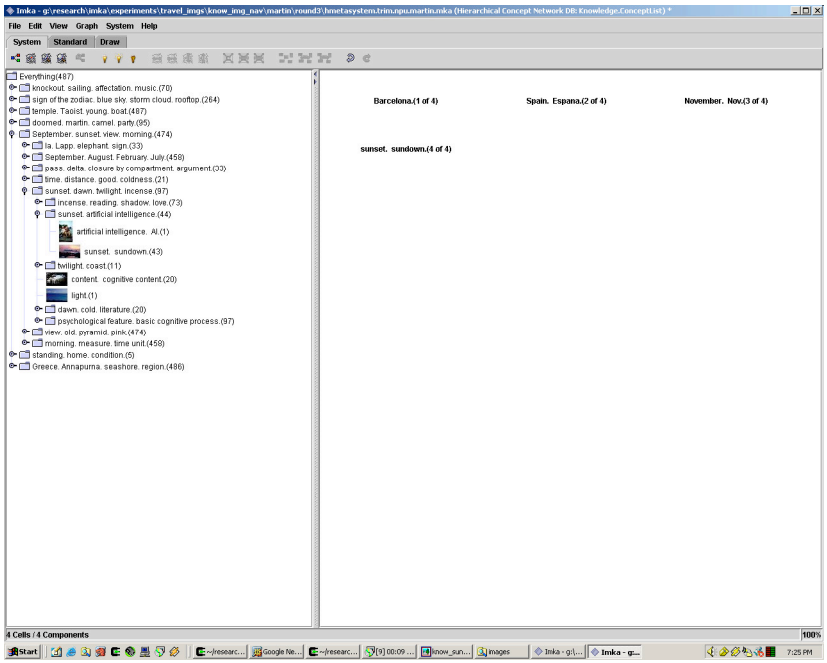


Figure 6.4: Screen shot of the MediaNet browser illustrating the visualization of the concepts to which an image is associated on the local view window (right side). An image can be associated with several concepts so no node is highlighted in the concept hierarchy on the global view (left side).

The MediaNet browser keeps a history of the concepts and images visited in navigating the medianet hierarchy. At any point of the browsing experience, users have the option of going back and forth to the previous and next screens of concepts and images already visited. This functionality is similar to the back/forward buttons in Windows Explorer and your favorite Web browser. Other basic functionality supported by the MediaNet browser is custom scaling for seeing the display larger or smaller, among others. Some of the browser functionality (e.g., show large resolution versions and annotations of images)

is easily configurable from pull-down menus, which facilitated evaluating and running different rounds of experiments in the user study as described in section 6.6.

6.4.2 Navigating Medianet Hierarchies

The proposed techniques for navigating medianet hierarchies follow the Visual Information-Seeking Mantra [128]: "Overview first, zoom and filter, then details-on-demand".

As described in section 6.3, the medianet at the lowest resolution level has n concepts. Each concept has m or fewer child concepts except for the leaf concepts, which are elementary perceptual or semantic concepts. The medianet hierarchy is a hierarchical structure; however, users can navigate the structure in a non-strictly hierarchical way. At any specific point in time, medianets of at most eight concepts are shown on the display. The reason for this is that human short-term memory has been shown to have a capacity of "the magical number seven plus or minus two" [96].

In the MediaNet browser, the users are first shown the medianet of n concepts at the top of the hierarchy, which acts as an overview of the entire collection of annotated images. The selected node is therefore the virtual root of the hierarchy (see Figure 6.1). Users can then select and zoom in any concept to view the part of the medianet corresponding to its child concepts at the higher resolution level. The zoom in command can also display concepts in the higher resolution level that are not child concepts but are similar, in terms of concept distances, to the child concepts. Therefore, the navigation of

the medianet hierarchy is not strictly hierarchical because zooming in a concept can display not only child concepts but also neighboring concepts. The inverse operation, the zoom out operation, is also supported and it selects and displays the child concepts of the parent concept of the currently selected concept at the lower resolution level.

Other concept navigation operations allow users to pan through the medianet at a given resolution level; these are the pan left/right operations. For these operations, the concepts at each resolution level are ordered in an array. The first concept is selected at random and put in the first position of the array. The most similar concept to the concepts already in the array is iteratively added to the array until all the concepts are sorted in the array. Concept similarity is measured as the number of shared neighbors between concepts. When users pan to the left/right concepts, the system selects and displays the next/previous concept in the array.

Users can request to see the images associated with any concept in the medianet hierarchy (see Figure 6.3). The leaves of the medianet hierarchy are elementary perceptual and semantic concepts. If the user selects and zooms in a leaf concept, the system directly shows the images associated with the concept. Concept occurrences in images are counted after propagation in the medianet as described in section 4.5.1. Users can also perform the inverse operation; in other words, they can request to see the concepts that occurred in selected images (see Figure 6.4). Users can then click on any of the concepts on display, which selects and displays the corresponding concept in the

medianet hierarchy. This is another non-strictly hierarchical navigation functionality of the MediaNet browser.

Images associated with a concept are shown as thumbnails on a 2D grid (see Figure 6.3). The system shows a maximum number of images at a time. Users can request the next or previous set of images for a concept, as applicable. The images are linearized so that similar images (in terms of feature distance) appear nearby. This is done by clustering the images and by serializing the images in each cluster as suggested by Craver et al. [38]. If users are interested in a particular image, they can view a large resolution version of the image together with the itemized annotations available for the image (e.g., title, description, place, and time of an image). In a similar way, the concepts associated with an image are shown on a 2D grid with a limited size (see Figure 6.4 and Figure 6.11). Users can request the next or previous set of concepts for an image, as applicable. The concepts for an image are ordered from least to most probable.

6.5 Visualizing Medianet Hierarchies

Users can browse a collection of annotated images by navigating the medianet hierarchy constructed from the annotated images, as described in section 6.4. The section describes the proposed techniques for visualizing hierarchies of medianets for users. We use fish-eye views (similar to [38], [50], [51], [148], etc.) for displaying concepts in medianets using representative text and images of the concepts. A medianet (or part of a medianet) is drawn on a 2D display using spring modeling [75]. The concept hierarchy is

displayed as a browsable index tree of text and images. Figure 6.5 illustrates the visualization of concepts, (parts of) medianets, and concept hierarchies.

6.5.1 Visualizing Concepts

Ideas from fish-eye views are used to provide information about the type and the probability of displayed concepts. Representative text and images for concepts are selected based on statistics and feature similarity. Figure 6.5, Figure 6.6, and Figure 6.7 illustrate the visualizations of several concepts.

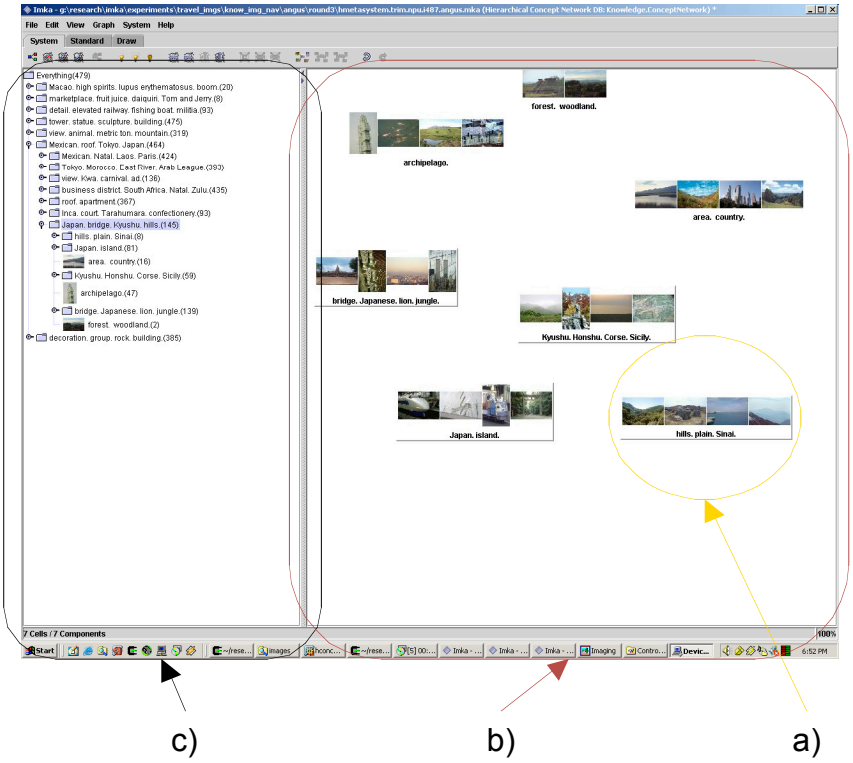


Figure 6.5: Screen shot of the MediaNet browser illustrating the visualization of a) concepts, b) (parts of) medianets, and c) concept hierarchies.

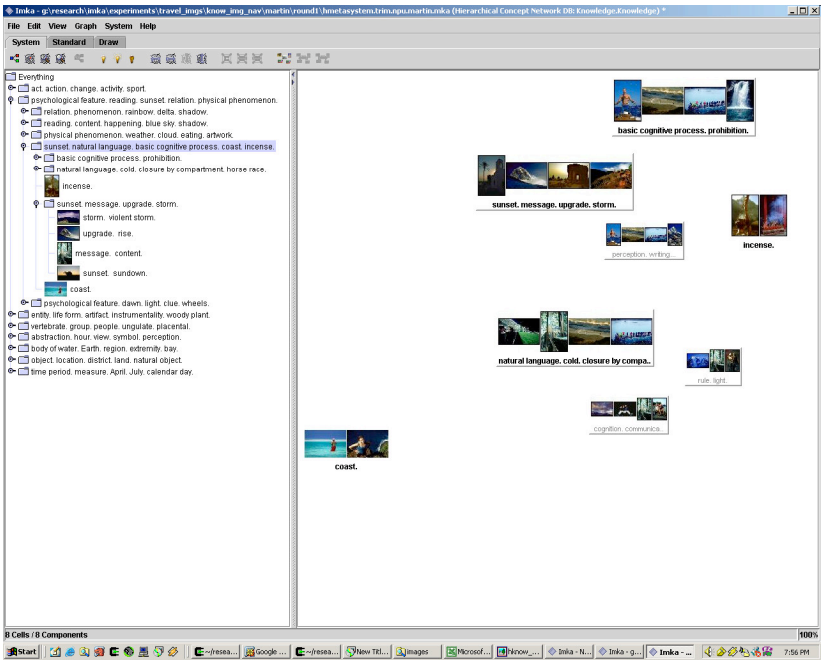


Figure 6.6: Screen shot of the MediaNet browser illustrating the visualization of neighbor concepts (e.g., "cognition"), super concepts (e.g., "sunset"), and elementary concepts (e.g., "coast").

The idea behind fish-eye views (similar to [38], [50], [51], [148], etc.) is to modify the attributes of the objects being displayed (e.g., size, orientation, and shape) based on the current viewing point. For example, an object close to the viewing point is in focus as opposed to an object that is far away. Therefore, we modify the attributes of concepts displayed on screen to indicate the type and the probability of concepts. Concepts with elevated border are concepts that are clusters of concepts in higher resolution levels. The size of concepts on screen can be modified to reflect the probability of concepts (as given by equation (6.1)); the larger the concept, the more probable (see Figure 6.7 for examples).

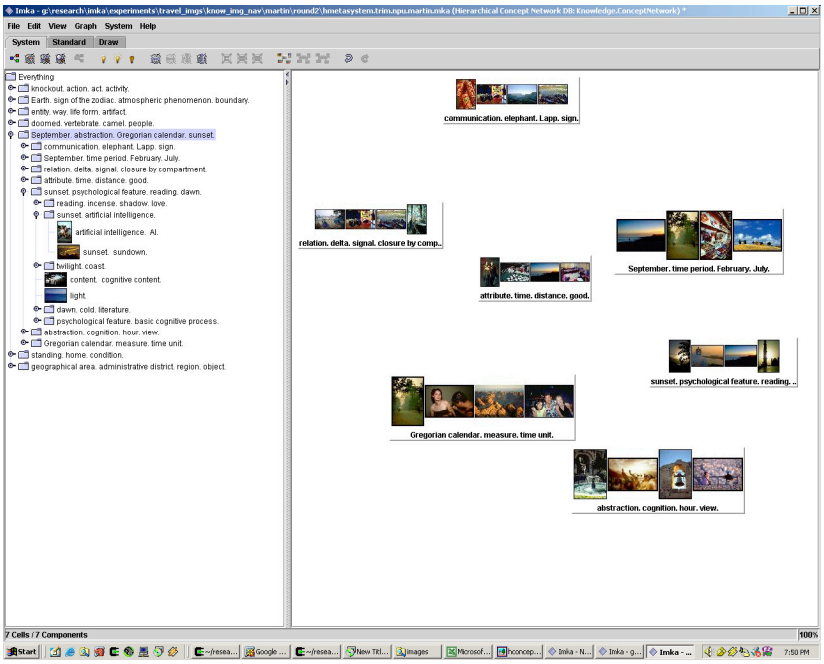


Figure 6.7: Screen shot of the MediaNet browser illustrating views of super concepts. The size of the view for each concept is scaled based on the number of associated images associated with the concept.

Concepts in medianets are displayed using their media examples, in our case, text and images. Two different views are created for each concept, a simple and an extended view. The way the views are created differs for each concept type, i.e., perceptual concepts (image clusters), semantic concepts (word senses), or super concepts (clusters of perceptual, semantic, and super concepts).

Perceptual concepts are image clusters based on visual and/or textual features (see section 4.3.2). The media examples of perceptual concepts are therefore the images grouped by the cluster. The simple view of a perceptual concept is the centroid of the cluster, in other words, the image closest to the center of the cluster based on the features used to create the cluster. In the extended view, four images are selected to

display the concept. These images are the centroids of clustering the images of the perceptual concept into four cluster using k-means. The clustering uses the same features used to construct the perceptual concept. If the perceptual concept has four or less images, all the images are used to visualize the concept.

Semantic concepts are word senses in WordNet (see section 4.4.2). The media examples of semantic concepts include the synonyms of the sense, as provided by WordNet. Other media examples of a semantic concept are the images on whose annotations the sense was detected. The simple view of a semantic concept is the first synonym of the sense together with the image with the highest concept occurrence after propagation. For example, the sense "plant, flora, plant life" is represented using "plant" and the image whose annotations have the highest number of occurrences of the sense "plant, flora, plant life". In a similar way, the extended view of a semantic concept consists of the two first synonyms of the sense (e.g., "plant, flora" in previous example) together with the four images with the highest concept occurrences. In the extended view, more synonyms and images help clarifying the intended meaning of concepts to users. If the semantic concept has four or less image examples, all the images are used to visualize the concept.

Super concepts are clusters of semantic and perceptual concepts, and other super concepts (see section 4.5.3). The simple view of a super concept is created using text and images in the simple view of the most frequent child concept. Child concepts are ordered in terms of concept probabilities as given by equation (6.1). The text in the

simple views of the four most common concepts is used to display the super concept. Four images are used to represent a super concept; these are the first four different images representing each of the four most frequent child concepts. If the super concept has four or less image examples, all the images are used to visualize the concept.

The simple and extended views are used for visualizing concepts in the MediaNet browser as follows. Extended views of semantic and perceptual concepts are shown in the local view window; whereas, simple views are shown for super concepts. Exceptions are concepts that are not children of the selected concept. These concepts are, instead, similar or neighbor concepts. Similar concepts are visualized using simple views. In addition, icons of similar or neighbor concepts are smaller and the text is painted in gray rather than black using again ideas from fish-eye views (see Figure 6.6 for examples).

6.5.2 Visualizing Medianets

A medianet (or part of a medianet) is drawn on s 2D display using spring modeling [75]. We position concepts in the (part of a) medianet on the display based on the distances of the relationships between concepts using the spring modeling algorithm.

There are many methods for drawing graphs (i.e., a set of nodes connected by arcs). The spring modeling algorithm [75] was selected because it has been found to minimizes arc crossings when drawing the graph [50]. This algorithm considers each pair of nodes in a graph to be connected by a virtual spring. The algorithm iteratively repositions the nodes of the graph so as to minimize the overall tension or energy of the system of springs.

Consider a string of negligible mass, one end of which is attached to a wall. The energy stored in the spring whose free end is stretched a distance X is given as follows:

$$E = \tfrac{1}{2} KX^2 \tag{6.3}$$

where K is the spring force. The total energy of the system of springs connecting the n nodes of a graph is given by:

$$E = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \tfrac{1}{2} k_{ij} (| v_i - v_j | - l_{ij})^2 \tag{6.4}$$

where v_i is the position of node i in the display, d_{ij} is the distance between nodes i and j, and l_{ij} and k_{ij} are the length and the strength of the spring between nodes i and j, respectively. l_{ij} and k_{ij} can be calculated in terms of the distance among nodes i and j –d_{ij}-, the length of a shortest side of the display area -L_o-, the minimum distance between two nodes– L', and the maximum distance between any two nodes, using

$$l_{ij} = \frac{L_0 \, d_{ij}}{\max_{z<l}(d_{zl})} + L' \text{ and} \tag{6.5}$$

$$k_{ij} = \frac{K}{d_{ij}^{\,2}}, \text{ respectively.} \tag{6.6}$$

Typical values used for K and L' are 1 and 50, respectively. The application of the spring modeling algorithm to draw (parts of) medianets is straightforward. The concepts are the nodes of the graph; and the relationships between the concepts, the arcs connecting the nodes. The distance of each relationship is calculated using equation (4.3). If two concepts are connected in the medianet by, at least, one relationship, the distance between the two concepts is the distance of the shortest relationship between them. If two concepts are not directly connected by a relationship, their distance is the maximum distance. The distance between the concepts is normalized between zero and one to obtain d_{ij} as follows:

$$d_{ij} = e^{-\frac{1}{(a(c+d))^{2b}}} \tag{6.7}$$

where a, b, and c are constants; typical values are 0.5, 0.5, and 1.5, respectively. To make the display of medianets less cluttered and confusing for non-expert users, relationships between concepts are usually omitted. As discussed below, a very short 5-minute instruction on the workings of the MediaNet browser was enough for non-expert users to use the MediaNet browser. We believe drawing the relationships among concepts is more important for expert users, which know the meaning and implication of different relationships.

6.5.3 Visualizing Concept Hierarchies

The hierarchy of medianets defines a hierarchy of concepts, which is displayed as a tree using representative text and images of the concepts.

The leaf concepts of the tree are represented using text and images; whereas the intermediate concepts of the tree are displayed using only text. Using images and text to represent all the concepts in the tree makes the display too cluttered, confusing, and complex. We display the perceptual and semantic concepts, i.e., leaf concepts of the tree, using the images in their simple views and the text in their extended views. We display super concepts in the tree using the text associated with their simple view. The procedure proposed for creating simple and extended views of concepts was described in section 6.5.1. The number of images associated with a concept can be attached to the text displayed for the concept in the index tree (see Figure 6.8).

6.6 User Study

In this section, we present the methodology and discuss the results of the experiments performed for evaluating the proposed techniques for image browsing. The main goal of browsing systems is to assist users in exploring unfamiliar collections. Thus, we evaluate the proposed techniques with a task-oriented, user-centered study. For baseline comparison we use the WordNet-based approach proposed by Yang et al. [165]. In particular, the purpose of the user study was to evaluate the subjective satisfaction, the effectiveness, and the efficiency of users in executing common browsing tasks using the

two image browsing systems. Evaluation measures were obtained by administrating questionnaires to subjects and by monitoring several parameters in the systems.

6.6.1 Methodology

In the user study, 26 subjects were asked to perform two main tasks using the two systems under evaluation on two image collections. The independent variable was the system type: the MediaNet or the WordNet browser. A diverse set of dependent variables indicative of subjective satisfaction and task execution (e.g., efficiency and effectiveness) were obtained through the administration of questionnaires and the monitoring of system parameters in the background (e.g. execution time), respectively. For the experiments, we employed two collections of about 500 travel photographs each taken by two different photographers. The main tasks were related to finding images for travel brochures about two specific topics simulating a real work task scenario.

To minimize the effects on the results of learning from one system, task, and image collection to the other, the order of the two systems, tasks, and image collections was rotated and assigned to subjects according to a Greco-Latin square design [162]. For example, a subject did task for topic A using system A on collection A, first, and, then, task for topic B using system B on collection B. The following subject performed task for topic B using system A on collection A, first, and task for topic A using system B and collection B, later, and so on so forth. Therefore, the user study was designed so that technical decisions in designing the experiments will have little or no effect on the results.

In this section, we describe in detail the methodology followed in the user study in terms of the image collections and their processing, the tasks, the systems, the hypothesis, the participants, and the measures.

6.6.1.1 Image Collections

We employ two collections of about 500 photographs taken by Angus McIntyre [90] and Martin Wierzbicki [161] in their travels around the world (see Figure 5.7.a and Figure 5.7.b), respectively. Both are amateur or almost professional photographers who have published their pictures online. We collected and picked these pictures because of our interest in the domain of consumer photographs: these pictures are similar to the ones taken by tourists on vacation. In addition, these pictures are annotated with short descriptions written by the photographers. Finally, the photographers travel to many of the same countries and continents.

We disambiguated the senses of words in the image annotations as described in section 4.4.2. The images were scaled down to a maximum height and width of 200 pixels. Color histogram was extracted from the scaled images. Words that appeared less than twice were discarded for the extraction of the word weighting features, whose dimensionality was further reduced to 125 bins using latent semantic indexing [42]. Images were clustered based on color histogram, log tf * entropy, and a concatenated color histogram + log tf * entropy feature vector, into 22 clusters for each feature. We use the clusters based on the concatenated color histogram + log tf * entropy feature for disambiguating

the word senses. This feature was selected because it provided the best word-sense disambiguation accuracy in a subset of 279 images whose annotations were manually labeled with the correct senses. The number of different disambiguated senses was 781 and 554 for the Angus and the Martin collections, respectively. It is important to point out that the medianets for both browsing systems did not include any perceptual concepts in this user study.

6.6.1.2 Tasks

We tried to place the subjects participating in the user study in a simulated work task scenario as in [72][153]. These scenarios seem to allow the information needs of subjects to evolve in the same dynamic manner as the needs of real users in real work situations. Subjects were asked to imagine they were travel agents and that one of their responsibilities was to design pamphlets including text and images for various travel destinations. Their main task was a search task that was defined as finding and selecting five relevant images for two specific pamphlets from two image collections with the help of two different image browsers. Figure 6.8 depicts the actual interface of the MediaNet browser as used by subjects during the user study.

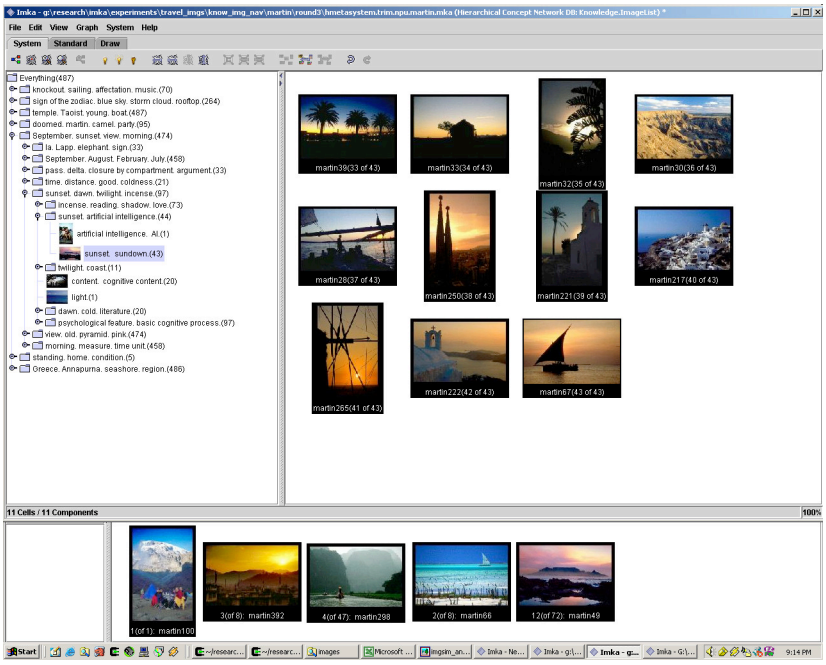


Figure 6.8: Screen shot of the interface of the MediaNet browser as used by subjects during the user study. The top part of the interface is the actual image browsing system; whereas users could drag-and-drop images relevant to the pamphlet on the bottom part.

As both photographers have visited similar places and taken similar photographs, it was not difficult to find travel destinations that could be satisfied using images from each collection. In particular, subjects were asked to find relevant images for pamphlets titled "Romantic Getaways" and "Buddhist Vacations" for which a brief textual description was provided. To minimize any bias in designing the tasks, the description of both pamphlets was taken from online web pages and later revised for a suitable length (paragraph) and for relevance to the travel photographs. We provide below the information about each pamphlet given to the subjects together with the URLs of the web pages from which the text was created.

- Pamphlet 1:
    - Title: "Romantic Getaways"
    - Text: " … Everyone has dreamed of that perfect romantic vacation, honeymoon, or getaway – warm tropical breezes, beautiful sunsets, or stunning mountain views – alone with your partner. Whether your dream includes an oceanfront cottage with strolls and picnics on the beach in the moonlight, or a mountain cabin with breathtaking sunsets on snow-capped peaks, you're sure to be inspired by these ideas. … "
    - URL: http://goflorida.about.com/cs/toppicks/a/aa012401a.htm
- Pamphlet 2:
    - Title: "Buddhist Vacations"
    - Text: " … Buddhism is a religion and philosophy based on the teachings of Gautama Siddhartha, who lived between approximately 563 and 483 BCE. This religion originated in India and gradually spread throughout Asia, to Tibet, Nepal, China, Vietnam, Laos, and Japan. Buddhist monks live a secluded life and meditate in temples where they seek the truth about existence. Traditionally, monks shave their heads, wear saffron robes, and have to beg for their food. … "

- URLs: http://en.wikipedia.org/wiki/Buddhism;
  http://www.kidcyber.com.au/topics/thaireligion.htm;
  http://www.personal.psu.edu/users/k/v/kvt103/kvt_cssb.html

In addition to the main search task of finding appropriate images for pamphlets, subjects were asked to perform overview and similarity tasks. After familiarization with each system and image collection for a few minutes, the overview task consisted on asking subjects to write down five words that better describe the content of the images in each collection and, then, to pick five words out of 25 choices with the same purpose. The similarity task was done once subjects finished the overview and search tasks for both systems. Subjects were shown six images. For each image, subjects were asked to select an image (out of two) that would likely appear in the same pamphlet with the given image. Subjects also had to provide a tentative title for the pamphlet. The purpose of the overview and the search tasks was to compare and evaluate the differences between the two browsers. However, the similarity task was aimed at studying how likely was for subjects to guess and agree on the similarity between images.

6.6.1.3 Systems

We evaluated the proposed methods for image browsing with the approach proposed by Yang et al. [165]. We refer to Yang et al.'s system as the WordNet browser.

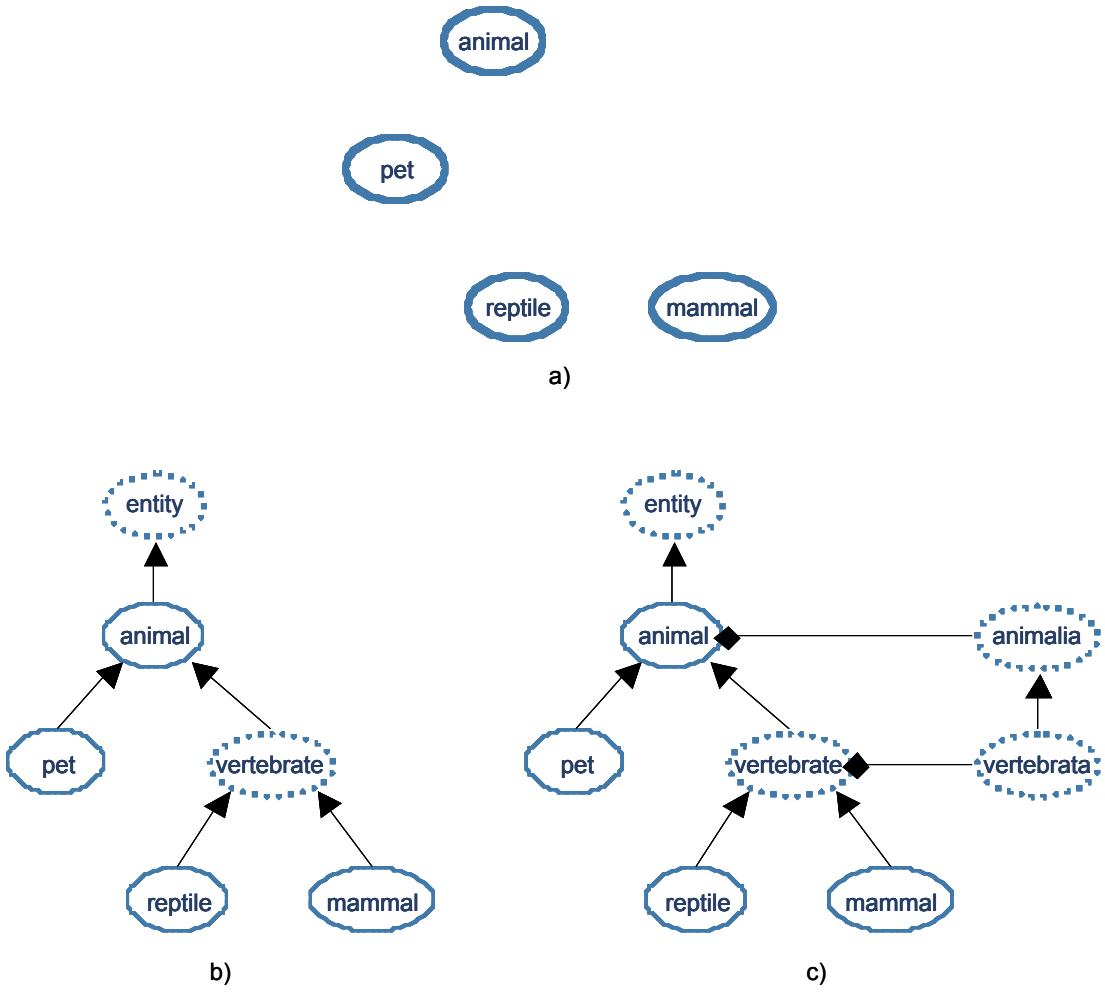


Figure 6.9: a) Set of disambiguated concepts with corresponding b) concept hierarchy in the WordNet browser, and c) medianet at the highest resolution level in the MediaNet browser. Solid circles represent originally disambiguated concepts; whereas, dash circles represent additional concepts taken from WordNet to build the concept hierarchy in both systems. The medianet hierarchy is generated by summarizing the medianet at the highest resolution level. UML conventions are used to represent the relationships between the concepts: an arrow means that that the origin concept is a specialization of the destination concept; the line finished in a diamond represents that the origin concept is part of the destination concept.

For each image collection, we built two browsing systems, the MediaNet browser and the WordNet browser, using the disambiguated senses of nouns in the image annotations. The MediaNet browser used the medianet of senses and relationships between senses (see section 4.4), and the summaries generated at different resolution levels (see section 6.3). The WordNet browser used instead the minimal concept hierarchy from WordNet including the disambiguated senses given by the specialization/generalization relationship. The hierarchy was however simplified to bypass senses with only one child. The concept hierarchy of the WordNet browser is thus a sub-network of the medianet at the highest resolution level in the MediaNet browser, which, in addition, contains additional concepts and relationships. Figure 6.9.a shows a set of disambiguated concepts whose corresponding concept hierarchy in the WordNet browser and the medianet at the highest resolution level in the MediaNet browser are depicted in Figure 6.9.b and Figure 6.9.c, respectively. Figure 6.10 shows the interface of the WordNet browser. The WordNet browser works in a similar way to the MediaNet browser. Some differences are that only text is used to represent each concept (a concept corresponds to one sense in WordNet); and that every level in the hierarchy is a set of concepts, not a concept network, as in the MediaNet browser.

The user study was divided into three rounds in which we changed the parameters and functionality of the two browsers. In particular, we could configure the systems for providing information about the number of images associated with each concept, for showing higher resolution views and annotations of images, and for displaying concepts

associated with each image. For the MediaNet browser, in addition, we could modify the number of similar and child concepts in the medianet hierarchy, the constants for calculating concept probabilities and obtaining representative words for concepts (see equation (6.1)), and the concepts shown to be related to selected images (e.g., all concepts in the medianet hierarchy, all leaf concepts, or the leaf concepts directly appearing in the image annotations). Table 6.1 lists the functionality and configuration of the WordNet browser and the MediaNet browser in each round. Both systems always provided the basic functionality of navigating the concept hierarchy and viewing the images associated with concepts.

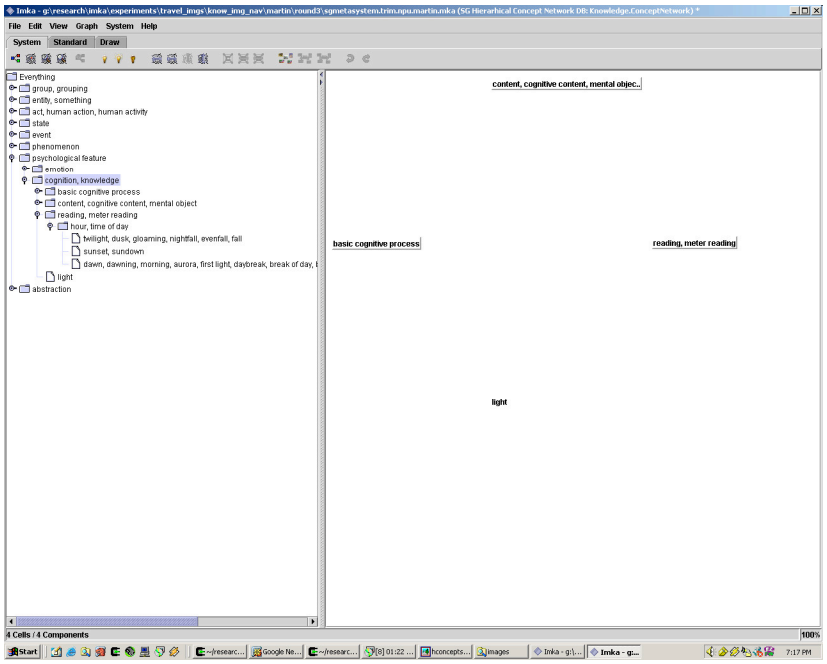


Figure 6.10: Screen shot of the WordNet browser illustrating the global view of the concept hierarchy on the left side and the local view of the concept visualization on the right side.

Table 6.1: Functionally and configuration of the WordNet browser and the MediaNet browser for each round. Both systems always provided the basic functionality of navigating the concept hierarchy and viewing the images associated with concepts. The table also includes the number of concepts and levels in the medianet or concept hierarchies for each browser. These numbers include the root node "Everything" that includes all the concepts at the lowest resolution level.

	WordNet Browser	**MediaNet Browser**
Round 1 **10 subjects**	View all concepts per image Angus: 1122 concepts, 11 levels Martin: 745 concepts, 12 levels	View all concepts per image At most five children per concept At most three neighbors per concept a = 1 in equation (6.1) Five keywords per concept Angus: 1503 concepts, 9 levels Martin: 1107 concepts, 8 levels
Round 2 **8 subjects**	(in addition to Round 1) Next/previous screens Number of images per concept	(in addition to Round 1) Next/previous screens Number of images per concept Concepts ordered by probability Images ordered by feature similarity At most seven children per concept No neighbors per concept a = 0.25 in equation (6.1) Four keywords per concept Angus: 1430 concepts, 7 levels Martin: 1037 concepts, 7 levels

Round 3 **8 subjects**	None (same concepts as levels as level 1)	(in addition to Round 2) Large and annotations per image a = 0 in equation (6.1) (same concepts as levels as level 2)

In the first round, both systems provided the functionality of navigating the concept hierarchy, viewing the images associated with any concept, and retrieving the concepts associated with an image. None of the systems could provide information about the number of images per concept, or high-resolution views and annotations of images. Regarding the concepts related to images, all the concepts in the concept hierarchy were shown, i.e., in the order of 20-50 concepts for the WordNet browser and of 75-150 concepts for the MediaNet browser (including super concepts in the medianets at different resolution levels). Figure 6.11 shows the visualization of the concepts associated with an image in the first round for the MediaNet browser. Images and concepts were not ordered in any specific way for display. The concept hierarchy in the MediaNet browser was constructed to have at most eight concepts at the lowest resolution level. From one resolution level to the higher one, concepts could have, at most, five children and three neighbors. Five words were selected to visualize each concept considering concept probabilities after propagation (a = 1 in equation (6.1)).

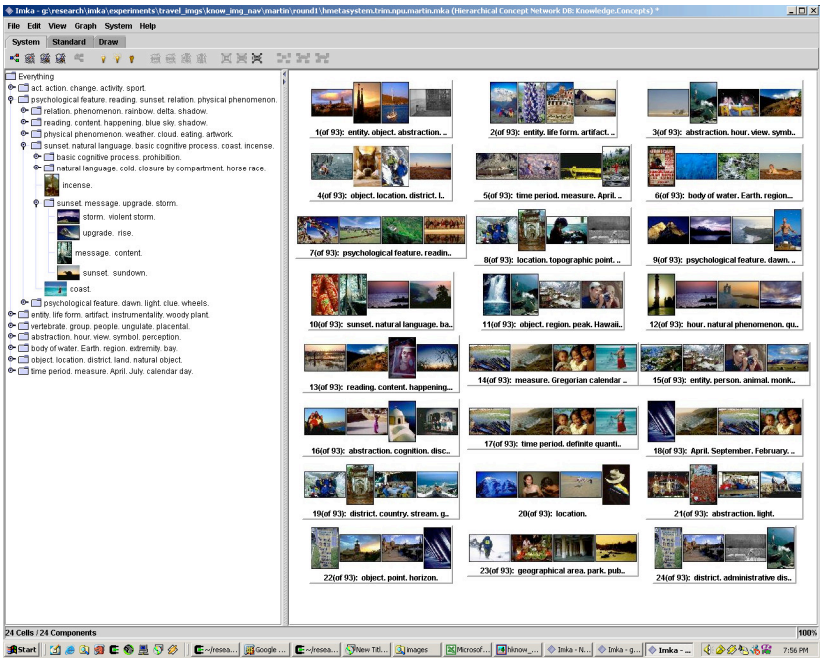


Figure 6.11: Screen shot of the MediaNet browser illustrating the visualization of the concepts associated with an image in the first round. Please, compare with Figure 6.4, which illustrates the visualization of the concepts associated with an image in the third round.

The two systems were modified based on comments received during the first round of the user study. In particular, functionality for going back and forward to the previous and next screen, and for showing the number of images associated with concepts was added to both systems. In the WordNet system, the number of images for each concept appeared after the concept's name in the hierarchy; whereas, the size of the concept icons on the MediaNet browser was scaled based on the number of images (see Figure 6.7). The larger the number of images; the larger the concept icon. The concept hierarchy in the MediaNet browser was constructed to have at most seven concepts at the lowest resolution level. Seven was also the maximum number of children of

concepts from one resolution level to the higher one. No neighbor concepts were used to navigate the medianet hierarchy in this round. Four words were selected to visualize each concept using the parameter a = 0.25 in equation (6.1). Only the concepts at the highest resolution level were returned for the images (~20-50 concepts per image; similar to the WordNet browser). Only text was used to represent these concepts. Images and concepts were ordered on screen based on feature similarity and concept probabilities, respectively.

The functionality of the two systems under evaluation was balanced in the first two rounds. The main functionality of the MediaNet browser was also included in the WordNet browser. In the third round, we compared the complete MediaNet browser with the basic WordNet browser as proposed by Yang et al. [165]. The basic WordNet browser only included the functionality for navigating the concept hierarchy and viewing the images associated with any concept. Therefore, users could not retrieve concepts associated with images; go back and forward to the previous and next screens; or view the number of images per concept. The complete MediaNet browser included the functionality of viewing high-resolution views and the annotations of images in addition to the functionality of the system in the second round. In the MediaNet browser, four words were selected to visualize each concept using the parameter a = 0 in equation (6.1). Only the concepts in the highest resolution level that appeared in the image annotations were returned for the images (~ 5-15 concepts per image). Figure 6.2, Figure

6.3, Figure 6.4, Figure 6.5, and Figure 6.10 are screen shots of the browsers in the third round.

6.6.1.4 Hypothesis

The main drawback of approaches that rely on single structures for organizing image collections such as the WordNet browser [165] is that the organization structure may become too large and complex. Then, non-expert users are likely to find the organization confusing, non-intuitive, and ineffective to browse. Examples of these approaches use subject hierarchies that are either manually created or based on existing thesaurus such as the WordNet browser.

Our image browsing methods try to address this drawback by summarizing the initial image organization structure at multiple resolutions levels, and by representing nodes in the summaries using multimedia, in our case, text and images. In addition, the functionality of the MediaNet browser has been tailored to user needs based on user feedback through this study including cross-referencing from images to concepts, going to next and previous screens of concepts and images, and displaying high-resolution views and annotations of images. Finally, we order images and concepts for display on screen based on feature similarity (see prior work on image multi-linearization [38] and image organization based on feature similarity [118]) and concept probability, respectively.

We would like to demonstrate the differences in the concept hierarchies between the two browsing systems in the different rounds of experiments. As an example, the path from the root of the concept hierarchy, "Everything", to the concept "sunset, sundown" for the Martin collection in each system was the following, each group of words between quotes represents one concept:

- WordNet browser (see Figure 6.10):
    - "Everything"
        - "psychological feature"
            - "cognition, knowledge"
                - "reading, meter reading"
                    - "hour, time of the day"
                        - "sunset, sundown"
- MediaNet browser, first round:
    - "Everything"
        - "psychological feature. reading. sunset. relation. physical phenomenon"

- "sunset. natural language. basic cognitive process. coast. incense"
    - "sunset. message. upgrade. storm"
        - "sunset. sundown"

- MediaNet browser, second round:
    - "Everything"
        - "September. abstraction. Gregorian calendar. sunset"
            - "sunset. psychological feature. reading. dawn"
                - "sunset. artificial intelligence"
                    - "sunset. sundown"

- MediaNet browser, third round (see Figure 6.2):
    - "Everything"
        - "September. sunset. view. morning"
            - "sunset. dawn. twilight. incense"
                - "sunset. artificial intelligence"
                    - "sunset. sundown"

Each concept in the paths above belongs to a different resolution level. To view any concept, the user would need to select and zoom in the previous concept in the path. This simple example shows that WordNet can be unintuitive and not very useful to common users in browsing or searching for images. Not many people would have looked for "sunset. sundown" under "psychological feature" or "cognition, knowledge", as an example.

The hypothesis of the user study is therefore that our image browsing approach results in better subjective satisfaction, efficiency, and effectiveness of users in performing common browsing tasks.

6.6.1.5 Subjects

We designed the scenario and the task of the user study for simplicity and generality so that non-specialized, non-expert people could participate in the study. Our sample of the user population consisted of 26 people with college or graduate education. The distribution of the subjects based on major was as follows: eight electrical engineers, eight computer scientists, two biologists, two psychologists, two business persons, one international affairs person, one mathematician, one broadcaster, and one teacher. The subjects were enlisted from the network of acquaintances of the main author of this thesis. All the subjects knew how to use computers but have no prior knowledge or experience in using any of the evaluated image browsers.

Only one subject participated in the user study at a time. In each occasion the procedure was as follows.

- an initial questionnaire,
- an introduction to the scenario of the search task,
- for each of the two systems in turn:
    - a demonstration of the system,
    - a training session on the system to get familiar with the system and the image collection,
    - an overview questionnaire in which the user selected relevant keywords for describing the image collection,
    - a text describing the written instructions for the search task,
    - a search session in which the user interacted with the system in pursuit of the search task, and
    - a search questionnaire,
- an image similarity questionnaire in which the user selected similar images to some given images and provided relevant keywords for the image pairs, and
- a final questionnaire.

The most important aspect of the study was to evaluate the differences between the WordNet and the MediaNet browsers through the results of the search task. The order of the last two bullets was exchanged in some questionnaires. A sample of the evaluation questionnaire used in the user study is included in appendix 11.2. Subjects were given unlimited time to finish the experiments for which they earned, in exchange, a gift certificate of $10.

6.6.1.6 Measures

The purpose of the user study was to evaluate the subjective satisfaction, the effectiveness, and the efficiency of users in executing common browsing tasks using the two systems. The questionnaires that users were asked to complete allowed us to collect their feedback regarding their satisfaction in using and executing the tasks with each system. In addition, several system parameters were tracked and monitored in the background to measure the effectiveness and the efficiency of both systems. In particular, the number of viewed images, the number and the type of executed browsing operations, and the time taken in each session, among others, were recorded without knowledge of the subjects. In addition, the images selected by the user to satisfy the requirements of the main search task were scored based on the instructions of the task.

6.6.2 Initial Questionnaire

The purpose of the initial questionnaire was to collect information about the participants' experience with computers and familiarity with image retrieval. This

questionnaire revealed that all the users knew how to use computers. In addition, only 31% of the subjects were familiar with image retrieval thus our sample of the user population was representative of common consumers.

6.6.3 Overview Questionnaire

In this section, we present the setup and discuss the results related to the overview questionnaire for studying the initial overview subjects got from browsing an image collection for a few minutes.

6.6.3.1 Setup

After a short demonstration of each system, subjects were asked to familiarize themselves with the system by freely browsing the image collection for a few minutes. Then, they were given a questionnaire to write and select descriptive words of the image collection. In average, each participant spent 4 minutes and 25 seconds, and used 54 browsing commands during the training session. Examples of browsing commands are zooming in/out, panning left/right, and retrieving the images-per-concept/concepts-per-image, among others.

Table 6.2: The 20 choices of key words. The "$p_o(c)$" and "p(c)" columns provide the probability of the concepts before and after propagation in the medianet at the highest resolution level, respectively. The highest values of each column are underlined.

	Angus Collection		**Martin Collection**	
	$p_o(c)$	**p(c)**	**$p_o(c)$**	**p(c)**
Asia	0.00	<u>3.67</u>	0.00	<u>14.70</u>
Balcony	0.16	0.16	0.00	0.00
Bridge	<u>0.82</u>	0.26	0.05	1.12
Bus	0.10	0.01	0.09	0.17
Country	0.10	<u>12.33</u>	0.00	<u>25.81</u>
Daiquiri	0.00	0.02	0.00	0.00
Dancer	0.02	0.22	0.05	0.09
England	0.13	0.74	0.00	0.02
Event	0.00	0.16	0.00	0.42
Heat	0.06	0.06	0.00	0.00
Home	0.00	0.90	0.05	0.09
Mama	0.06	0.06	0.00	0.00
Mountain Peak	0.14	0.46	<u>0.26</u>	<u>6.91</u>
Ocean	0.06	0.89	0.00	0.00
Person	0.28	<u>11.11</u>	0.00	<u>7.10</u>
River	<u>0.38</u>	<u>3.97</u>	<u>0.76</u>	<u>7.53</u>
Road	<u>0.28</u>	<u>1.20</u>	<u>0.26</u>	0.88
Temple	<u>1.42</u>	<u>1.65</u>	<u>1.01</u>	1.68
Tourist	0.03	0.03	0.00	0.00
Wind	<u>0.32</u>	0.44	<u>0.15</u>	0.19

After the training session with each system and image collection, subjects were asked to write down five words (free words) that better describe the content of the images in each collection and, then, to pick five words out of 20 choices with the same purpose (key words). The 20 choices of key words are shown in Table 6.2; they were the same for both image collections although they were ordered differently on the questionnaires. We generated the list of 20 choices by combining some of the most and the least frequent concepts for each image collection.

6.6.3.2 Results

Table 6.3 and Table 6.4 show the nine most frequently selected free words and key words, respectively, by users in each image collection. The table also lists the probabilities of the words in the selections by subjects, and in each collection before and after propagation in the medianet at the highest resolution level (i.e., $p_o(c)$ and $p(c)$, respectively). Based on these probabilities, we decide the text to be displayed for each super concept as described in section 6.5.1. For each participant, we added the probabilities of the selected free words or key words for each pair image collection and browsing system.

The mean probabilities of the word selections in the corresponding image collection for each system before (a=0) and after (a=1) propagation in the highest resolution medianet are shown in Figure 6.12. As the Gaussian distribution was found to be a good fit for the distribution of the mean word probabilities for each system and each subject, we used a

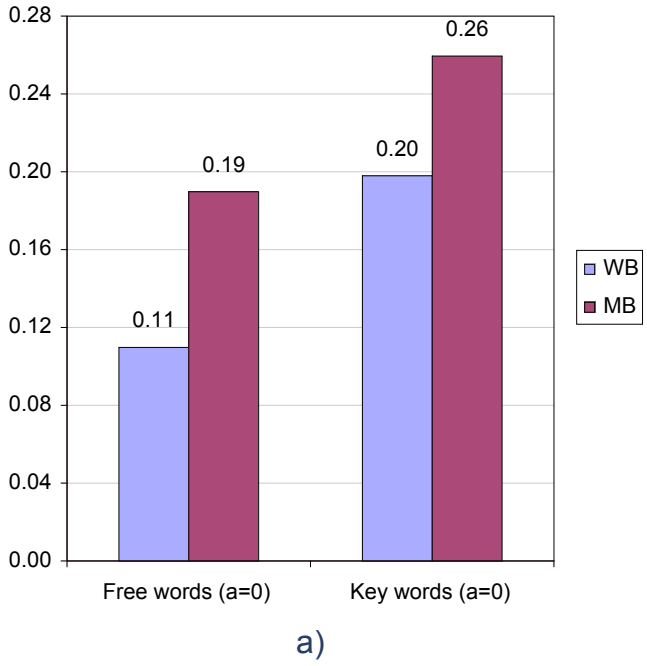
parametric technique, the Annova 2 test, to analyze the significance of the variance among these results. Both the mean differences for free words with after-propagation probabilities ("Free words (a=1)") considering all rounds, and with before-propagation probabilities ("Free words (a=0)") only considering rounds two and three are significant at a level of $p < 0.05$ ($p=0.0472$ and $p=0.0483$, respectively), i.e., with a 95% confidence interval.

Table 6.3: Most frequent free words specified by the subjects for the two image collections. The "Subjects" column indicates the word probability among the free words specified by all the users. The "$p_o(c)$" and "$p(c)$" columns provide the probability of the concepts before and after propagation in the medianet at the highest resolution level, respectively.

Angus Collection				**Martin Collection**			
Words	**Subjects**	**$p_o(c)$**	**$p(c)$**	**Words**	**Subjects**	**$p_o(c)$**	**$p(c)$**
Person	5.38	0.28	11.11	Person	13.85	0.00	7.10
Travel	5.38	0.00	2.21	Landscape	6.92	0.00	0.00
Animal	5.38	0.54	6.41	Travel	6.15	0.00	0.85
Nature	4.62	0.00	1.26	Nature	5.38	0.00	0.92
Building	3.85	0.95	5.89	Animal	4.62	0.00	3.02
Culture	3.08	0.00	0.00	Culture	3.08	0.00	0.00
Place	2.31	0.63	2.57	Colorful	3.08	0.00	0.09
City	2.31	0.19	14.31	Country	2.31	0.00	25.81
Object	2.31	0.00	68.22	Asia	2.31	0.00	14.70

Table 6.4: Most frequent key words selected by the subjects for the two image collections. The "Subjects" column indicates the word probability among the key words selected by all the users. The "$p_o(c)$" and "$p(c)$" columns provide the probability of the concepts before and after propagation in the medianet at the highest resolution level, respectively.

Angus Collection				**Martin Collection**			
Words	**Subjects**	**$p_o(c)$**	**$p(c)$**	**Words**	**Subjects**	**$p_o(c)$**	**$p(c)$**
Country	13.85	0.10	12.33	Person	16.15	0.00	7.10
Person	10.77	0.28	11.11	Tourist	14.62	0.00	0.00
Tourist	9.23	0.03	0.03	Asia	11.54	0.00	14.70
Event	5.38	0.00	0.16	Mountain Peak	11.54	0.26	0.6
Road	5.38	0.28	1.20	Event	6.92	0.00	0.43
Heat	2.31	0.06	0.06	Ocean	6.92	0.00	0.00
Bridge	2.31	0.82	0.26	Temple	6.15	1.09	1.68
Wind	0.77	0.32	0.44	Heat	3.85	0.00	0.00
Bus	0.77	0.10	0.09	River	3.85	0.76	0.88



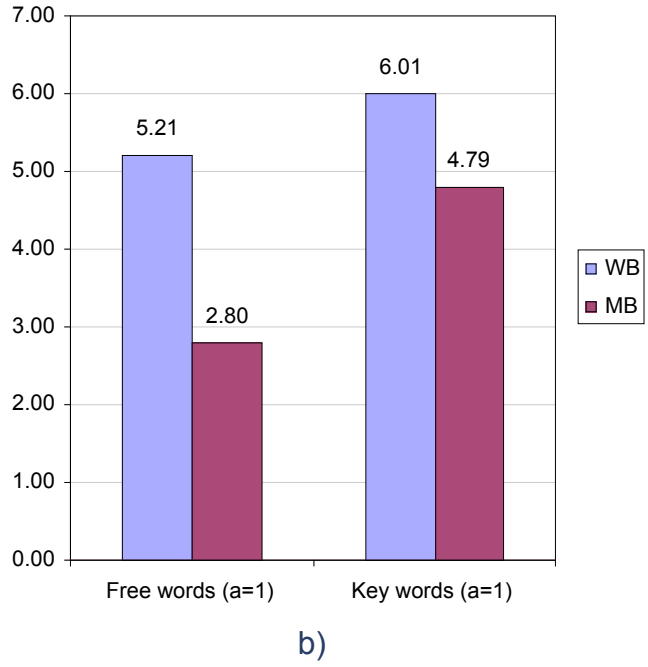


Figure 6.12: Mean probability (in percentages) of the free words and key words selected by subjects using the WordNet browser (WB) and the MediaNet browser (MB) in the corresponding image collection considering a)

concept probabilities before propagation (i.e., a=0 and $p_o(c)$ in equation (6.1)), and b) concept probabilities after propagation (i.e., a=1 and p(c) in equation (6.1)) in the medianet at the highest resolution level. The results come from all 26 subjects.

6.6.3.3 Discussion

Through this task, we are investigating if the word selections from users match the distribution of the concepts in the annotations before or after propagation in the concept network. Overall, the WordNet browser provided overviews of the image collections that consisted of the more frequent concepts after propagation (a = 1); whereas, the overviews from using the MediaNet browser match more closely the distribution of concepts in the image annotations before propagation (a = 0). These differences are statistically significant.

In any case, the overviews provided by the WordNet browser are static and can only match the concept distribution after propagation. However, the overviews provided by the MediaNet browser can be adapted to user preferences and needs to represent the concepts before and after propagation by modifying the parameter a in equation (6.1): values close to 1 for after propagation distribution; values close to 0 for before propagation distribution.

6.6.4 Search Questionnaire

In this section, we present the setup and discuss the results related to the search questionnaire for evaluating the subjective satisfaction in executing the search task.

During the search task, subjects had to find and select five relevant images for a given travel pamphlet.

6.6.4.1 Setup

After each search session on one of the systems, users were asked to complete a questionnaire about the task they were given, the browsing system they used, the feelings they had, the images they selected, and the success they had performing the task.

Each subject was asked to indicate, by making a selection from a 7-point Likert scale, the degree to which they agreed or disagreed with 16 statements about various aspects of their experience of using each system (for the actual statements see appendix 11.2). The statements focused on five different aspects:

- three statements focused on the given task (clear, simple, familiar);
- four of the statements focused on the browsing system (easy, intuitive, useful, stimulating);
- two focused on the user's feelings in interacting with the system (in control, comfortable);
- six focused on the images in the subjects' mind or selection (initial idea, matched initial idea, changed in mind, alternate ideas through browsing, happy with selected images or precision, saw all applicable images or recall); and

- the last one focused on the degree of success in performing the task.

Please, note that we use the terms "precision" and "recall" to refer to the statements asking subjects how happy they were with the selected images, and whether they felt they had seen all applicable images for the search task. These definitions are subjective; and, therefore, slightly different from the standard definitions of the precision and recall measures for evaluating the effectiveness of retrieval systems [71].

6.6.4.2 Results

Each user was asked to respond to the 16 statements twice, after each search task on the two different systems. The result was a set of 26 x 2 x 16 = 832 scores on a scale of 1 to 7 (with 1 representing the response "I strongly agree" and 7, "I strongly disagree"): 26 subjects scoring each of two systems with respect to each of the 16 statements. The mean scores for the five aspects and all the 16 statements are shown in Figure 6.13 and Figure 6.14, respectively. As the Gaussian distribution was not a good fit for the scores, we used a non-parametric technique, the Friedman test, to analyze the significance of the variance among the results. The subset of results with statistical significance of at least $p <= 0.05$ are given in Table 6.5 together with their means. Please, note the user study was designed so that technical decisions in designing the experiments will have little or no effect on the results (see section 6.6.1).

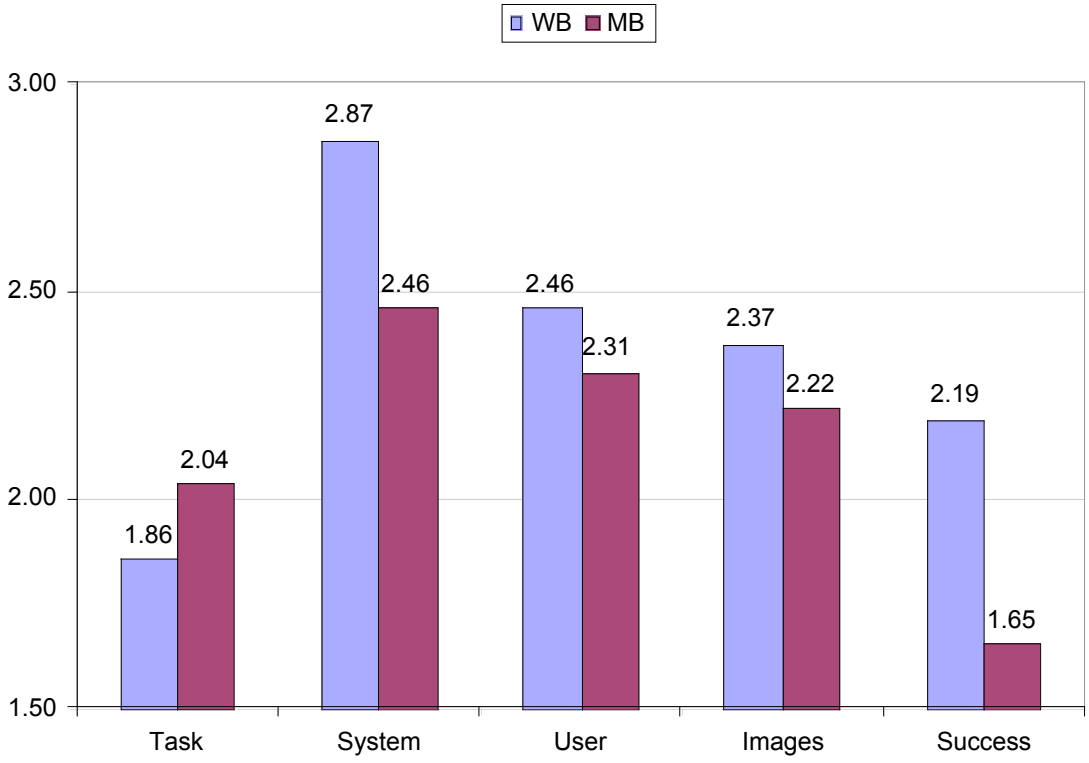


Figure 6.13: Mean scores for the WordNet browser (WB) and the MediaNet browser (MB) per evaluated aspect (i.e., task, system, user, images, and task success) in the search questionnaire (value range 1-7, lower = better). The results come from all 26 subjects.

6.6.4.3 Discussion

The trend is for the scores for the WordNet browser to be poorer than the MediaNet browser's. Therefore, overall the MediaNet browser outperformed the WordNet browser. This supports our initial claim that our proposed techniques result in better subjective satisfaction than prior work. The most significant differences between the two systems are found when comparing the statements related to the system and the success in executing the search task. The variance analysis points out the advantage of the

MediaNet browser as being significantly more useful, easy, and stimulating than the WordNet browser. In addition, the results reflect that the MediaNet browser helps users perform search tasks with more success. There were no significant differences for statements concerning the task, which shows that the tasks were balanced.

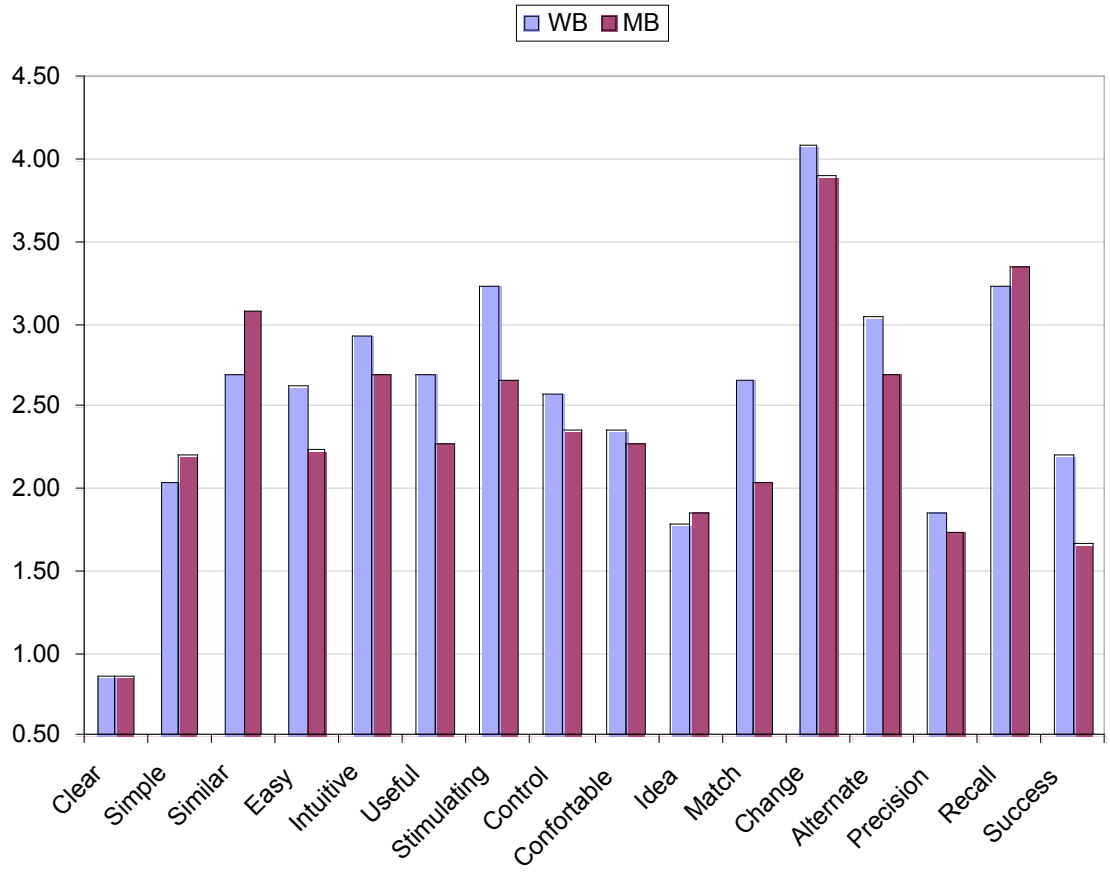


Figure 6.14: Mean scores for the WordNet browser (WB) and the MediaNet browser (MB) per each of the 16 statements (value range 1-7, lower = better) in the search questionnaire. The results come from all 26 subjects.

Table 6.5: Mean scores and p-values for the statistically significant result differences between the scores of the MediaNet and the WordNet browsers to the search questionnaire. The Friedman test was used to analyze the variance among the results.

All Rounds				
Type	**Statement**	**WB**	**MB**	**p-value**
System	Useful	2.69	2.27	0.0201
	All	11.46	9.85	0.0495
Success	Success	2.19	1.65	0.0348
System + User + Images + Success	All	33	30	0.0412
Rounds 2 and 3				
Type	**Statement**	**WB**	**MB**	**p-value**
System	Easy	3.06	2.44	0.0348
	Stimulating	3.56	2.56	0.0027
	All	12.63	10.06	0.0039
Round 3				
Type	**Statement**	**WB**	**MB**	**p-value**
User	Control	2.5	1.88	0.0455
	All	5.38	4.12	0.0455
Images	Idea	2.38	3.00	0.0455

We consider the results in the second and third rounds of the user study more significant for evaluating and comparing both systems. In the first round of experiments, users consistently requested the same desirable functionality for the systems (e.g., go back and forward to the previous and next screen, and information about the number of images associated with concepts) most of which was added to the systems in the second round.

In addition, we modified the value of the parameter "a" in equation (6.1) to fit user preferences and feedback. The first round was therefore like a training session to configure the functionality and parameters of the systems for the following rounds. The extra functionality was removed from the WordNet browser in the third round to investigate the most important functionality based on the users' point of view and needs.

The functionality that was pointed out as most useful and essential for an image browsing system and, in particular, the MediaNet browser, was: the multimedia visualization of concepts; the cross-referencing from images to concepts; high-resolution views and annotations for images; and the navigation buttons for going to the next and previous screens. In addition, subjects requested better names for the concepts and search functionality. One important limitation of both systems that subjects usually commented on was image misclassifications, i.e., finding images in non-related concepts. The main cause of these mistakes was errors in the automatic disambiguation of the senses of words in the image annotations (see section 4.4.2).

6.6.5 Image Similarity Questionnaire

In this section, we present the setup and discuss the results related to the image similarity questionnaire for evaluating the subjects' perception of similarity among images.

6.6.5.1 Setup

The purpose of this questionnaire was not to compare the two systems but rather to investigate the perception of image similarity by subjects in terms of semantic concepts. For this questionnaire, the first eight subjects were shown Figure 6.15; whereas; Figure 6.16 was shown to the remaining subjects. For each image on the left, subjects were asked to pick the image on the right that would most likely appear in a travel pamphlet together with the image on the left. In the topic column, subjects were asked to fill out the possible title of the pamphlet. Invariably, one of the image choices had a common semantic concept with the given image on the left. The other image was not similar from the given image based on their concepts' distances in the medianet (see section 4.5.1). The purpose of the similarity task was to study how often subjects could agree and guess the two similar images with the common semantic concept in the extracted medianet.

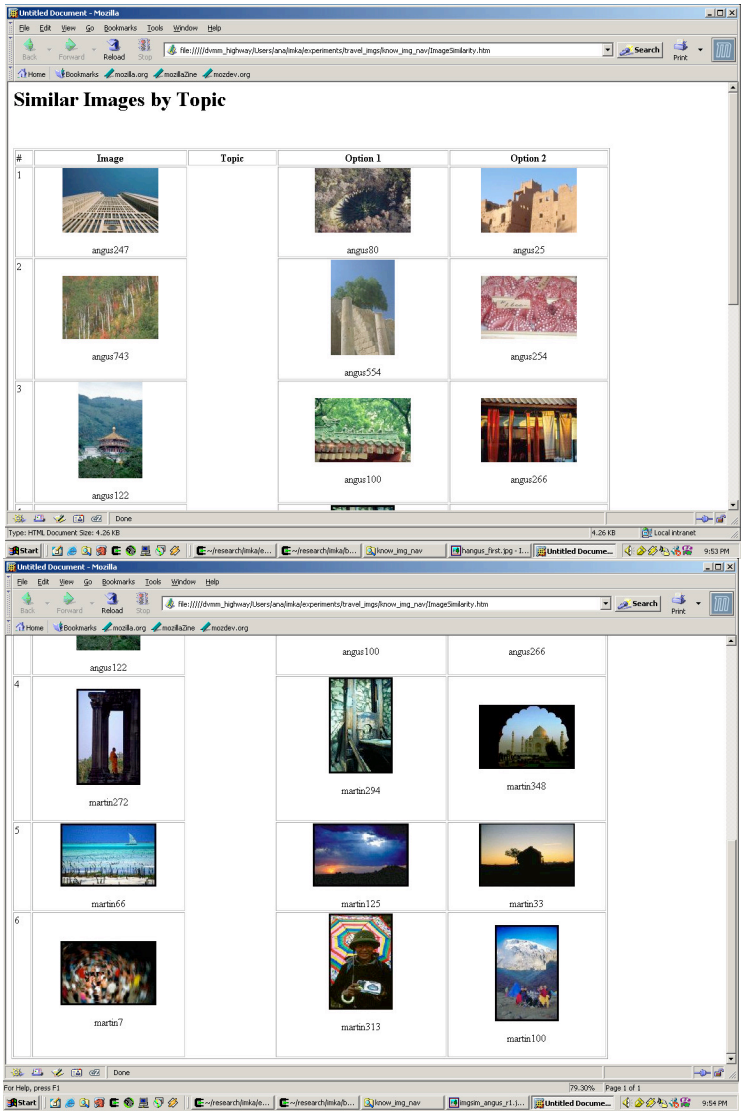


Figure 6.15: Images shown for the image similarity questionnaire from the Angus collection (top) and the Martin Collection (bottom) in the first round.

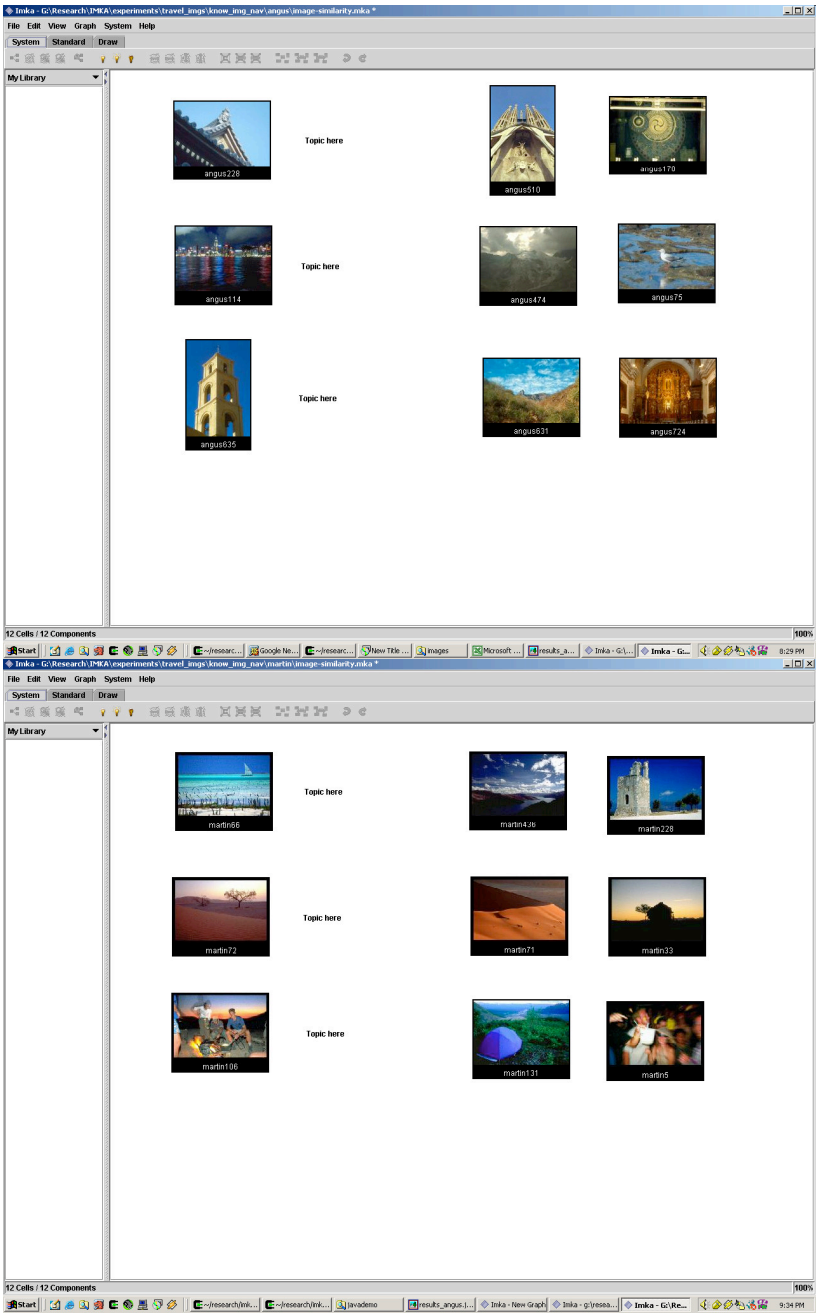


Figure 6.16: Images shown for the image similarity questionnaire from the Angus collection (top) and the Martin Collection (bottom) in the second and third rounds.

6.6.5.2 Results

To analyze the user choices, we assigned scores to the selected images and the specified titles. A score of one was given to each similar image that was selected. In addition, 0.5 was added to the score if the title of the pamphlet was the common concept between the images. Figure 6.17 shows the histogram of image similarity score values for the subjects. Therefore, image similarity scores for a subject could vary between 0 and 9. In other words, the y axis represents the number of subjects who got the specific value of image similarity score in the x axis.

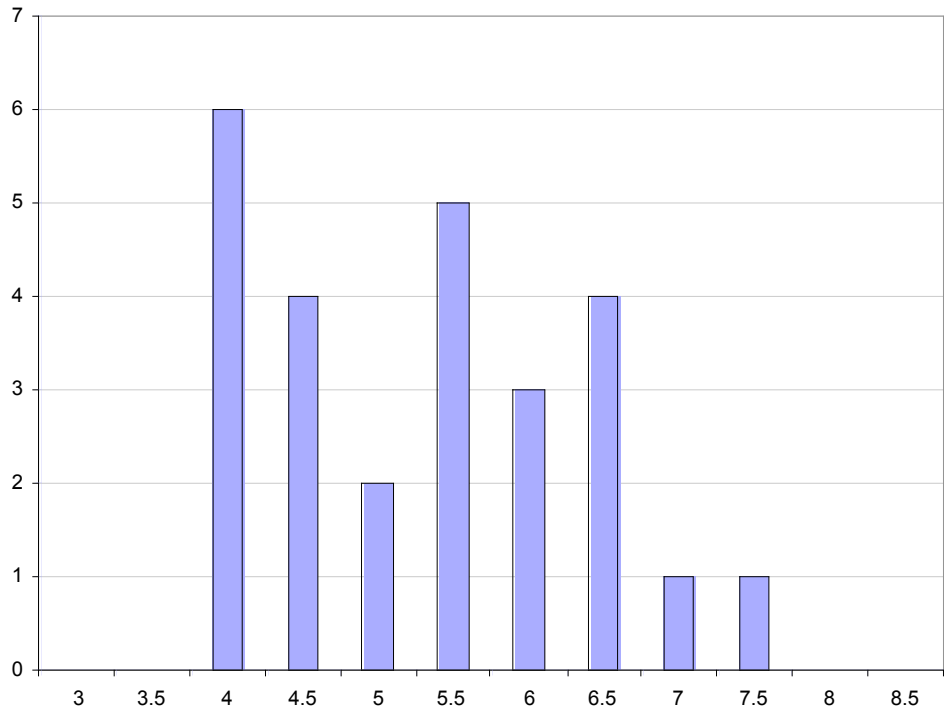


Figure 6.17: Histogram (number of subjects; y axis) of image similarity score values (value range 0-9, higher = better; x axis). The results come from all 26 subjects.

6.6.5.3 Discussion

As can be seen from the Figure 6.17, most of the subjects obtained at least 50% of the maximum score for image similarity. In other words, most subjects predicted the similar images and the common concepts with an accuracy of about 50%. However, it is important to point out the large variability between the scores across users, which demonstrates human subjectivity in judging in terms of image similarity.

Interestingly enough, several users comment on the difficulty of this task. An explanation for this could be that humans are not used to executing these kinds of tasks consciously. Most of the image similarity processing in humans is done in a subconscious way. Another reason could be the wide range of possible answers to the posed questions in terms, for example, of the similarity criteria between images.

6.6.6 Final Questionnaire

In this section, we present the setup and discuss the results related to the final questionnaire for comparing and commenting on the two browsing systems.

6.6.6.1 Setup

After completing all the tasks and questionnaires, subjects could comment on the two systems. This feedback of this questionnaire was used to modify and add additional functionality to the systems from one round to the next (see section 6.6.1.3 and Table 6.1). Eight subjects (the ones in the third round) also ranked the two systems in order of

preference with respect to (i) the one that was most useful in the execution of the task, and (ii) the one that they liked best.

6.6.6.2 Results

The mean ranks of the two systems based on usefulness and likeness are shown in Table 6.6. The significance of the mean differences was calculated using the Friedman test because the distribution of the ranks was not found to fit closely the Gaussian distribution. The value of the Friedman test was found to be statistically significant at a level of $p < 0.01$ ($p=0.0047$) for the ranks on the best-liked system.

Table 6.6: Mean ranks of the WordNet browser and the MediaNet browser in terms of the most useful and the best liked system. P-values are provided for significant mean ranks between the two systems. The Friedman test was used to analyze the variance among the results. The results come from the eight subjects in the third round.

Rank	WB	MB	p-value
Most Useful	1.75	1.25	-
Best Liked	2.00	1.00	0.0047

6.6.6.3 Discussion

The results in Table 6.6 show that the ranks for the two systems were sampled from different populations with statistical significance for the best liked system. Our conclusion, therefore, is that subjects liked the MediaNet browser significantly better

than the WordNet browser. Subjects also found the MediaNet browser more useful for the search task we set them.

Respondents who ranked the WordNet browser as the most useful system appreciated the system's specialization/generalization organization and naming of the concept hierarchy; however, they often recognized the difficulties in navigating such a structure. For example, a user said that "The WordNet browser is more organized …" and "The naming of the WordNet browser is more appropriate but starts with concepts that are too abstract (e.g., entity) and one has to descend step by step (tedious and slow)". Also referring to the WordNet browser, other users mentioned "… the organization of the labels was not very intuitive… but … was however more useful" and "… was easier to find all images that satisfied the conditions …".

On the other hand, the responses of people who preferred the MediaNet browser showed the advantages of the multimedia concept visualization, the cross-referencing between concepts and images, and the large resolution versions and annotations of images, among others. These browsing capabilities made the browsing system, for example, more stimulating, more useful, and easier to use. A user pointed out that "The MediaNet browser was easier to use. What I thought most useful was that I could get information or related categories of the images … Also, the MediaNet browser seems to have a better naming/labeling/organizing folders. I liked the thumbnail images …". Another user found the MediaNet browser most useful and liked it best because he/she "… could view what was in each folder and also enlarge photos and read explanation for

each image …". Another user pointed out that the MediaNet browser "… is more useful (cross-referencing). It is very useful to see all the images of a concept and all the topics of each image. An image without its description is not very useful … and … there has to be zoom for the images!".

6.6.7 System Monitoring

In this section, we present the setup and discuss the results related to the monitoring of system parameters for measuring the effectiveness and the efficiency of the browsing systems.

6.6.7.1 Setup

In addition to the use of questionnaires, several system parameters were monitored and tracked in the background while the subjects executed the tasks. This was done to measure the effectiveness and the efficiency of the browsing systems. In particular, the number of viewed images (all and unique images), the number and the type of executed browsing operations, and the time taken for each session (not including the time to fill out the questionnaires), among others, were recorded without knowledge of the subjects. In addition, the images selected by the user to satisfy the requirements of the browsing task were scored based on the instructions of each task.

6.6.7.2 Results

For each browsing system, the mean number of all viewed images together with the mean number of unique viewed images is shown in Figure 6.18. The mean length in minutes and the mean number of browsing operations executed during each session are shown in Figure 6.19 and Figure 6.20, respectively. The distribution of these values was found to be close to the Gaussian distribution so we used parametric techniques to analyze the variance among the results, the Annova 2 test. The subset of results with statistical significance of at least $p <= 0.05$ are given in Table 6.7 together with their means.

The images selected by subjects for each pamphlet (search task) were scored based on the text provided for the pamphlet (see section 6.6.1.2). The reason for this is that the title and the description were the only information provided to the subjects for each pamphlet. An image was assigned one point for each object mentioned in the text that was depicted in the image (e.g., sunset and mountains for the romantic pamphlet, and monks and temples for the Buddhist pamphlet). Table 6.8 lists all scoring objects for each pamphlet. When an object was given a score of one, more generic objects were not counted. For example, in the Buddhist pamphlet, a picture taken in China would get one point for China but no points for Asia. Each image was awarded at most three points. We believe this is a good measure because it is not subjective to the impressions of any human judge. The mean image scores for the WordNet browser and the MediaNet browser were 9.50 and 10.77, respectively (value range = 0-15, higher = better). The

Friedman test demonstrated the statistical significance of the differences between the two means at a level of $p < 0.05$ ($p = 0.0499$). The Friedman test was used because the data was found to have a non-Gaussian distribution.

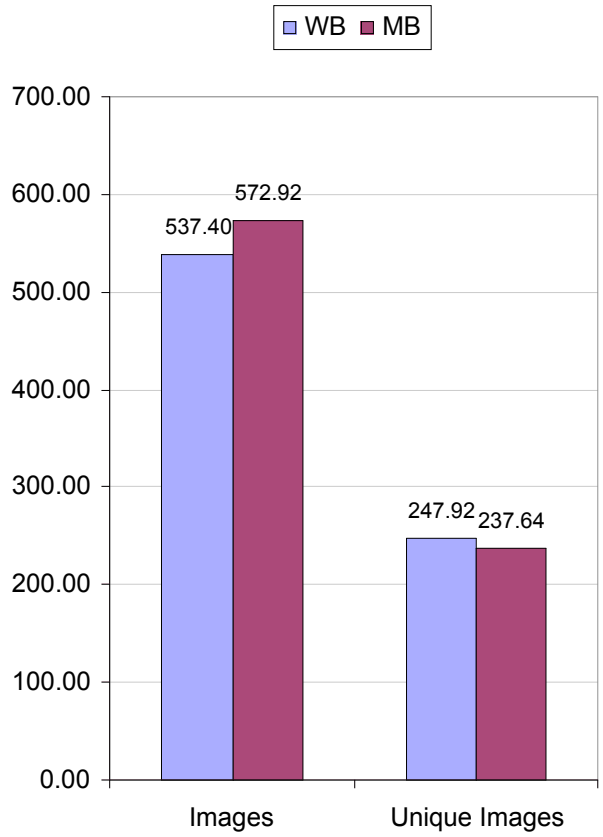


Figure 6.18: Mean values for the WordNet browser (WB) and the MediaNet browser (MB) of the number of images and number of unique images seen by the subjects during the experiments. The results come from all 26 subjects.

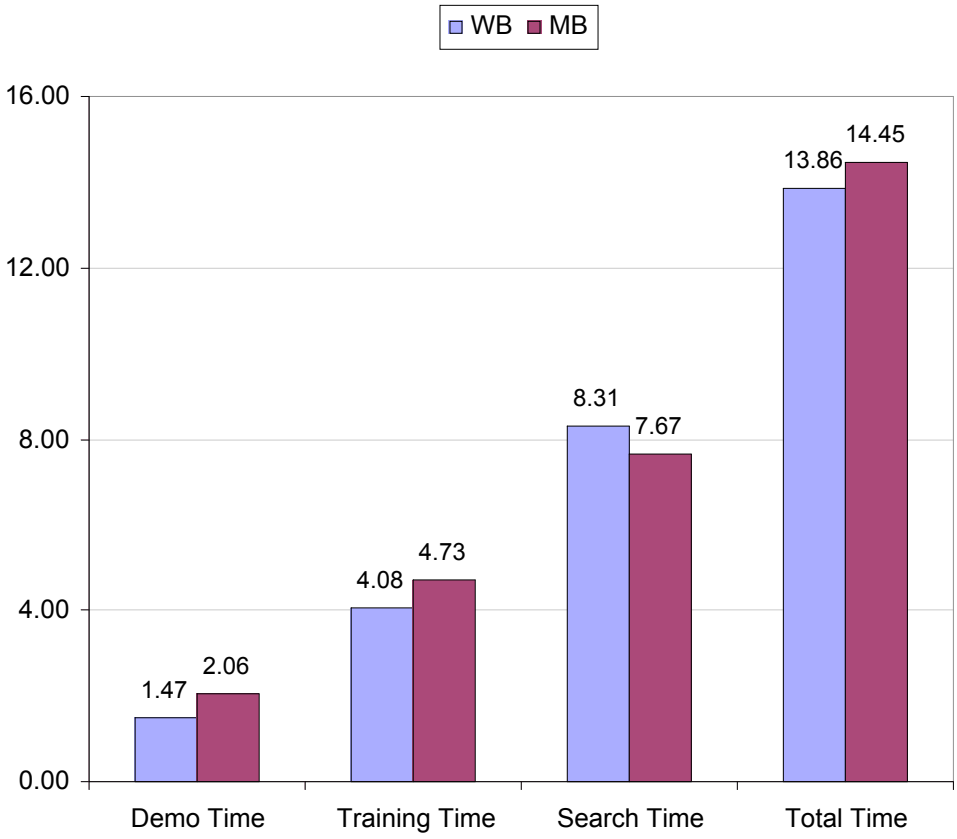


Figure 6.19: Mean values for the WordNet browser (WB) and the MediaNet browser (MB) of the duration (in minutes) of the demonstration, training, search, and all sessions. The results come from all 26 subjects.

6.6.7.3 Discussion

The MediaNet browser was more efficient than the WordNet browser in performing the search task: the number of unique viewed images, the search time, and the number of browsing operations were smaller. The difference in the number of executed browsing operations was statistically significant.

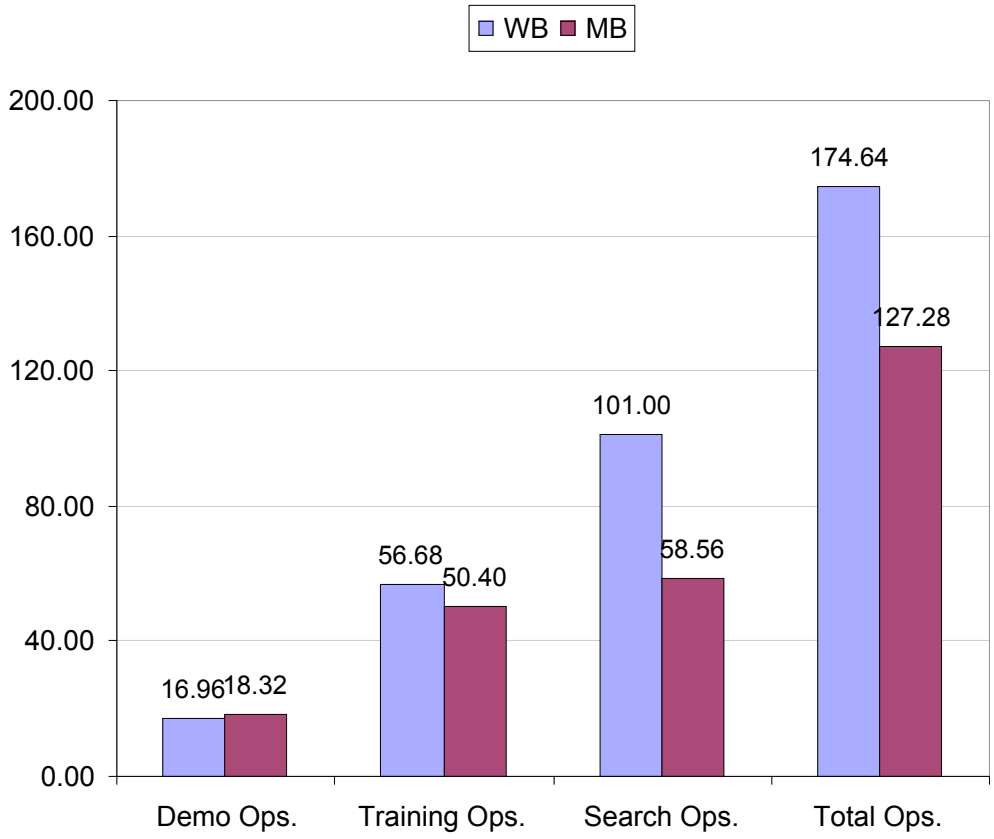


Figure 6.20: Mean values for the WordNet browser (WB) and the MediaNet browser (MB) of the number of browsing operations executed during the demonstration, training, search, and all sessions. The results come from all 26 subjects.

Table 6.7: Mean scores and p-values for the statistically significant result differences between the monitored system parameters of the MediaNet and the WordNet browsers. The Anova 2 test was used to analyze the variance among the results.

All Rounds				
Session	**Measure**	**WB**	**MB**	**p-value**
Search	Number of browsing operations	101.00	58.56	0.0100
All	Number of browsing operations	174.64	127.28	0.0099

Table 6.8: Scoring objects for the romantic and the Buddhist pamphlets.

Romantic Pamphlet		**Buddhist Pamphlet**	
Cabin	Mountain	Temple	Laos
Beach	Tropical	Monk	China
Snow	Ocean	Asia	Vietnam
Sunset	Couple	India	Buddha
Moonlight	Vacation	Tibet, Nepal	Saffron robes

In terms of effectiveness, the MediaNet browser was significantly better than the WordNet browser. The image scores about the relevance of the selected images for each pamphlet were significantly higher for the MediaNet browser. The MediaNet browser can therefore help users find more relevant images using fewer resources (e.g., time and browsing operations).

6.6.8 Discussion

In an extensive task-oriented, user-centered study, we evaluated the proposed approaches for image browsing, the MediaNet browser, with respect the browsing system proposed by Yang et al. [165], the one we call the WordNet browser. We have demonstrated the superiority of the MediaNet browser in terms of subjective satisfaction, efficiency, and effectiveness.

In the user study, 26 subjects were asked to perform two search tasks using the two systems on two image collections. The independent variable was the browsing system type. A diverse set of dependent variables indicative of subjective satisfaction, efficiency, and effectiveness were obtained through the administration of questionnaires to users and the monitoring of system parameters in the background (e.g. execution time), respectively.

Overall the MediaNet browser outperformed the WordNet browser. The MediaNet browser results in better subjective satisfaction as shown by analyzing the answers to the questionnaires. The most significant differences between the two systems are found when comparing system aspects (e.g., usefulness, easiness, and stimulation) and the success in executing the search task. The MediaNet browser was more efficient than the WordNet browser in performing the search task in terms of fewer unique viewed images, less search time, and significantly fewer browsing operations. The effectiveness of the MediaNet browser was demonstrated by more relevant images selected for each search task or travel pamphlet.

The functionality that was pointed out as most useful and essential for the MediaNet browser was: the multimedia visualization of concepts; the cross-referencing from images to concepts; high-resolution views and annotations for images; and the navigation buttons for going to the next and previous screens. In addition, subjects requested better names for the concepts and search functionality. We believe this functionality together with the method for constructing the concept hierarchy are the causes for the superior

performance of the MediaNet browser (see above). One important limitation of both systems that subjects usually commented on was image misclassifications, i.e., finding images in non-related concepts.

6.7 Summary

This chapter proposes novel techniques for automatically organizing and browsing annotated images based on extracted and summarized knowledge in the form of hierarchies of medianets.

Annotated images are organized in medianet hierarchies (i.e., concepts network hierarchies), which are constructed based on knowledge extracted and summarized from images and annotations. The initial medianet discovered from the collection is clustered hierarchically resulting in a medianet hierarchy. The discovered medianet can include perceptual knowledge (e.g., image clusters and relationships), semantic knowledge (e.g., word sense and relationships), and statistical interrelations among these. Users can then browse the annotated images by navigating the resulting medianet hierarchy. Ideas from fish-eye views and spring modeling are exploited for displaying concepts using text and images, and for drawing networks, respectively. In spite of the medianet hierarchy, its navigation does not necessarily have to be strictly hierarchical.

An extensive task-oriented, user-centered study has demonstrated the superiority of the proposed techniques in terms of subjective satisfaction (e.g., more useful, easier, more stimulating, and more successful), efficiency (e.g., fewer viewed images and executed

browsing operations), and effectiveness (e.g., better images selected for two travel pamphlets) of users in performing common browsing tasks such as searching for images related to a specific topic. We have compared the performance our approach with the thesaurus-based image browsing system proposed by Yang et al. [165] through the administration of questionnaires to subjects and the monitoring of system parameters in the background (e.g. execution time), respectively.

The user study has also pointed out the subjectivity (1) in the annotation of images (see section 6.6.3) and (2) in the appreciation of image similarity (see section 6.6.5). As future work, we are interested in extending the MediaNet browser with controls and relevance feedback mechanisms so that the medianet hierarchy and its visualization can be personalized and adapted to individual users' preferences. As mentioned above, the interface of the MediaNet browser is flexible enough to capture sophisticated feedback from users to modify, personalize, and optimize the medianet hierarchies (e.g., drag images and concepts around, and delete, paste, and copy nodes from one screen to another). The sophisticated user feedback could be used to modify, personalize, and optimize the medianet hierarchies in a much more powerful way than related systems that only allow users to modify clusters [33][62][7][81] or layouts [125] of images. Another interesting direction for future work is to extend the user-centered, task-oriented design and evaluation of to other multimedia information systems.

7 Knowledge-Based Image Retrieval

7.1 Introduction

This chapter focuses on novel methods for retrieving images based on medianets constructed from annotated images using WordNet. Current approaches in image retrieval lack flexibility; they are often constrained to features extracted from images and/or their annotations. In chapter 4, we presented our techniques for automatically discovering medianets from annotated images in the form of image clusters, words senses, and relationships among them.

Users need and want tools for effectively and efficiently retrieving multimedia, preferably, at the semantic level (e.g., based on people and objects in multimedia). However, many current multimedia retrieval systems use low-level features failing to meet user needs. For example, the study presented in [73] found that less than 20% of the attributes used by humans in describing images for retrieval were related to visual features. In addition, Armitage and Enser [1] found that most users want to retrieve images depicting specific objects, phenomena, or events, among others.

This chapter focuses on the retrieval of images using a medianet built from annotated images and the electronic dictionary WordNet. Most current image retrieval approaches

are often based on visual features (e.g., color and texture) failing to meet the user needs at the semantic level. Other approaches try to propagate annotations or labels to images without annotations and to retrieve images based on a combination of extracted textual and visual features. Therefore, these approaches are limited to the processing of the images, and its annotations or labels; they do not exploit contextual relations among concepts such as those provided in WordNet and the MediaNet framework.

In this chapter, we present innovative approaches towards image retrieval based on extracted knowledge, and visual and textual features [22][23]. The main contribution of this work is the enhancement of typical Content-Based Image Retrieval (CBIR) systems with medianets and advanced query processors. Figure 7.1 shows the components of the proposed retrieval system, which we refer to using the term "MediaNet retrieval system". The medianet is separated from the collection from which images are retrieved and can include both semantic and perceptual concepts. Medianets were constructed from annotated images following a process slightly different from the one described in chapter 4. The main difference is that the senses of the words in the annotations were disambiguated semi-automatically using WordNet [95] and human input. The query processor translates, expands, and/or refines incoming queries across modalities, as needed, using the medianet. The transformed queries are then inputted to a typical CBIR system and the results merged by the query processor. Initial experiments have demonstrated the superior retrieval effectiveness of the MediaNet retrieval system compared to the typical CBIR system for a semantic query. The proposed approaches

offer potential for enhancing image retrieval; however, due to a time limitation, this thesis has only explored and evaluated two particular case scenarios.

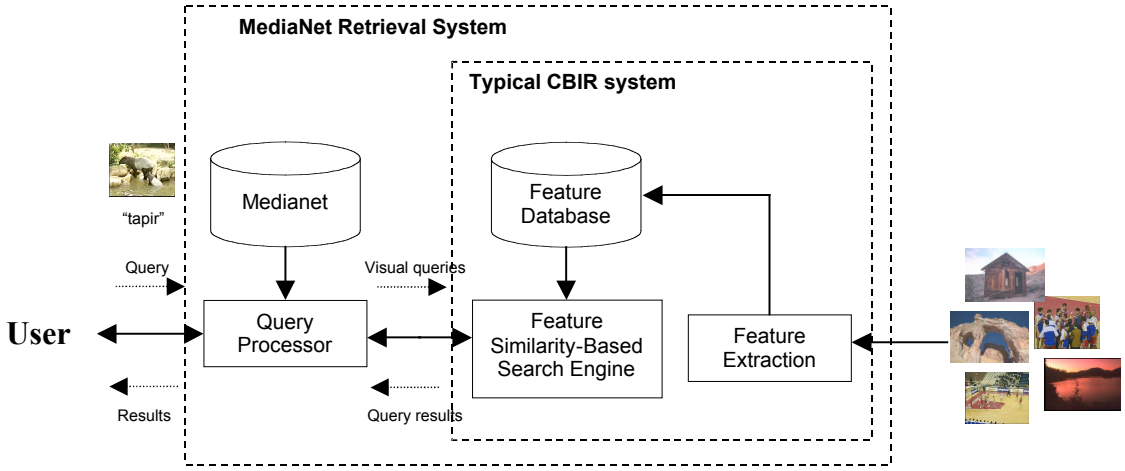


Figure 7.1: Components of the MediaNet retrieval system. MediaNet retrieval system extends a typical content-based retrieval system with a medianet and a query processor. A typical CBIR system is composed of feature extraction, feature database, and search engine components.

7.1.1 Outline of the Chapter

The rest of the chapter is organized as follows. In section 7.2, we review some related work on image retrieval. In section 7.3, we present the construction of the medianets from annotated images. We explain the way images are retrieved using the medianets in section 7.4. In section 7.5, we present the experiments performed for evaluating the proposed image retrieval methods. Finally, we conclude with a summary of the chapter and a discussion of future work in section 7.6.

7.2 Related Work

Relevant prior work on multimedia retrieval in the framework of multimedia information systems was presented in section 1.2. These approaches are limited to the processing of the images, and its annotations or labels; they do not exploit external knowledge resources such as WordNet. The most promising approaches for advanced multimedia retrieval systems seem to be the ones that combine information retrieval, content-based retrieval, and image classification techniques [92][144][145]. These approaches work with images with annotations or labels. Concept networks or hierarchies are built for a specific domain. Then, classifiers are built for automatically propagating the annotations or labels to images without annotations. Incoming queries in any of the supported media are expanded, translated, and refined across different media, as necessary (e.g., keyword query is translated into visual query). There is little work in the literature about this kind of approaches; an example is the Multimedia Thesaurus [144][145]. However, the Multimedia Thesaurus assumes annotations or labels are available for most images in the collection and it is limited to semantic concepts.

Our image retrieval framework enhances typical CBIR systems with medianets and advanced query processors (see Figure 7.1). The medianet is separated from the collection from which images are retrieved and can include both semantic and perceptual concepts. The query processor processes the queries across different media and merges the results for the user.

7.3 Constructing Medianets

For this work and experiments, medianets are constructed semi-automatically using annotated images and the electronic dictionary WordNet as shown in Figure 7.2. This process is slightly different from the one described in chapter 4 as this work preceded the development of the medianet extraction techniques presented in that chapter. The main difference is that the word-sense disambiguation is done semi-automatically under human supervision.

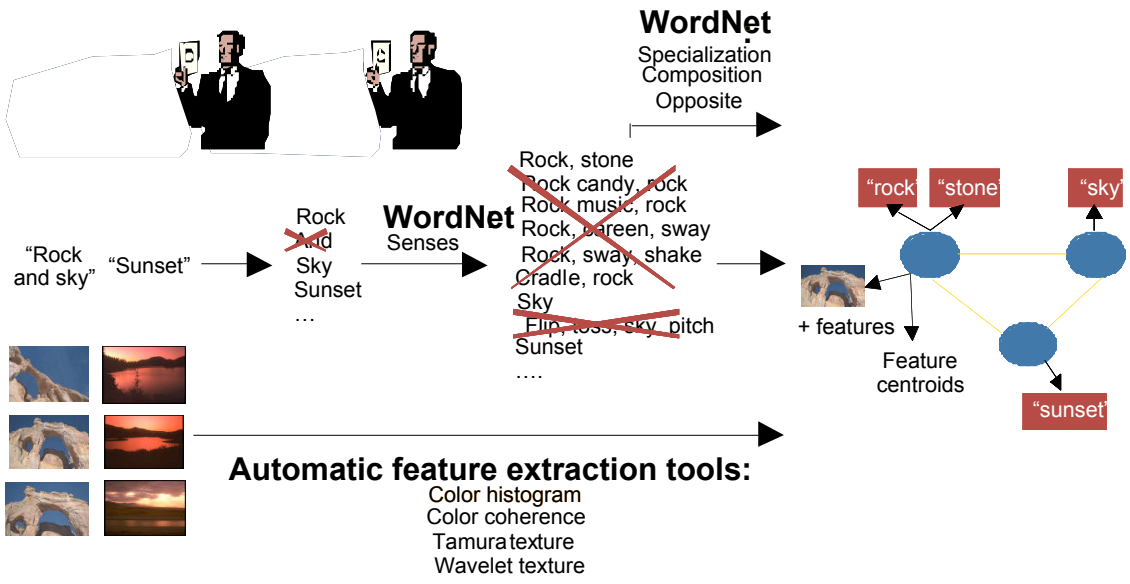


Figure 7.2: Procedure followed for semi-automatically constructing a medianet from a collection of annotated images.

First, stop words are removed from the annotations. All the possible senses of the remaining words are found in WordNet. Then, human supervisors decide the correct sense(s) of each word based on the corresponding image and/or annotations of the

word. As an example, for an image of a "rock, stone", the human supervisor discarded senses "rock candy, rock", "rock music, rock", and "cradle, rock", among others (see Figure 7.2). For several words, the human supervisor had to look at both the image and the annotations of a word to decide its correct sense. The reason for this is that image annotations are sometimes too short to provide enough context to disambiguate word senses. This highlights the importance and the difficulty of disambiguating image annotations for which we propose a novel automatic method that combines text and image data in section 4.4.2.

A medianet is then built from the disambiguated senses and their relationships. Opposite, specialization, and composition relationships among the disambiguated senses are, then, found in WordNet. Each sense is considered a concept in the medianet. The relationships between the detected senses in WordNet are the relationships between the corresponding concepts in the medianet. The final step is to extract visual features from the images associated with each concept. We find the centroid of each concept by averaging the features vectors of all the images associated with the concept. The feature centroid of a concept is used to detect new images that may be related to the concept as described in the next section. Perceptual knowledge and knowledge summaries could also be generated as described in sections 4.3 and 4.5, respectively; however, it was not done for the experiments described in section 7.5.

7.4 Retrieving Images Using Medianets

We propose to enhance typical CBIR systems with medianets and advanced query processors. The components of the MediaNet retrieval system are shown in Figure 7.1. It is important to note that the underlying CBIR system is a typical feature similarity-based image search engine composed of feature extraction, feature database, and search engine components. In this section, we introduce the MediaNet retrieval system and describe the procedure for processing queries using medianets.

7.4.1 The MediaNet Retrieval System

Figure 7.1 shows the components of the MediaNet retrieval system. In addition to the CBIR system (i.e., feature database, feature extraction, and search engine modules), the MediaNet retrieval system includes a medianet and an advanced query processor.

The medianet is separated from the collection from which images are retrieved and can include both semantic and perceptual concepts, as described in section 7.3. Queries from users can be words and/or images, which are translated, expanded, and/or refined by the query processor, as needed, using the medianet. The transformed queries are then inputted to a typical CBIR system. A user query at the query processor can result in the submission of several visual queries to the typical CBIR system. The query processor also merges the results returned by the typical CBIR system for each user query.

7.4.2 Processing Queries

The query processor uses the medianet to process incoming textual or visual queries from users, and to merge the results from the typical CBIR system. The procedure is shown in Figure 7.3.

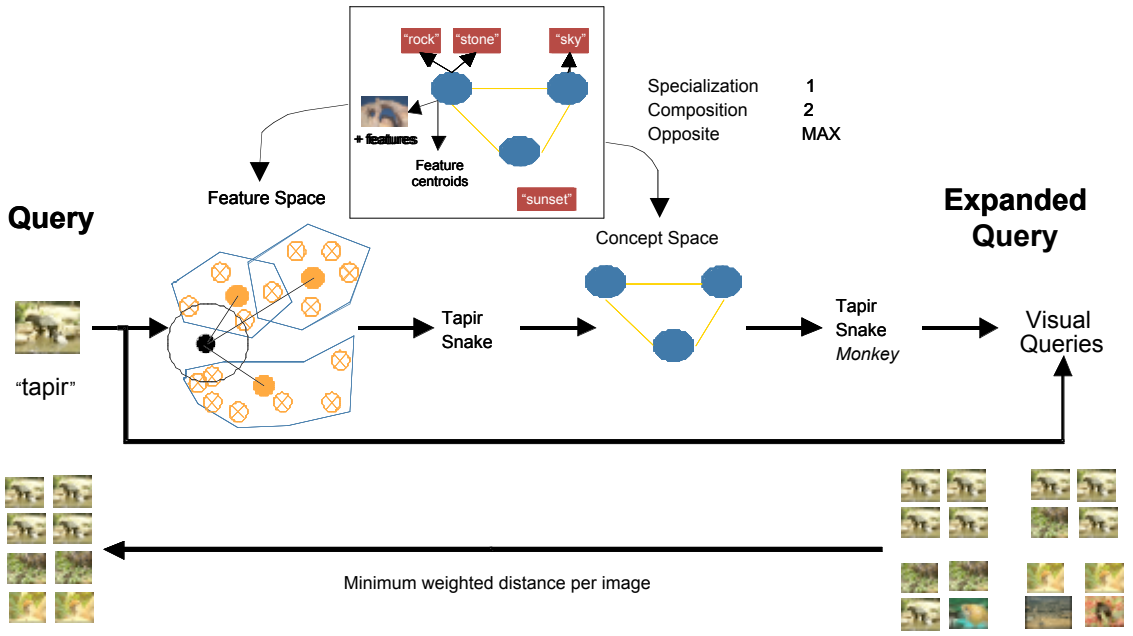


Figure 7.3: Multimodal query translation, expansion, and refinement procedure at the query processor.

First, the query processor classifies incoming queries, either words or images, into relevant semantic and perceptual concepts using concept detectors. The initial set of detected concepts is extended with semantically and perceptually similar concepts. Finally, images are retrieved that match the final set of concepts and their associated words and images. The retrieved images are ordered based on how closely they match

the input query and the concept set. Higher importance is assigned to the initial set of detected concepts than the additional concepts.

For image queries, relevant concepts are found by measuring the distance between visual features extracted from the query images and the feature centroids of the concepts in the medianet. The lower the distance, the more relevant the concept. If the queries are words, the concepts corresponding to the senses of the words are selected. The initial set of detected concepts is extended with other similar concepts. The distance between concepts is calculated by assigning weights to each relationship type and by finding the length of the shortest distance path between the concepts. Examples of weights for the specialization, composition, and opposite relationships are shown in Figure 7.3 (i.e., weights of 1, 2, and maximum, respectively). The relationship weights are static in the current implementation; however, the weights could be adapted based on user feedback and concept distances obtained as described in section 4.5.1. Each concept in the final relevant concept set is assigned a score based on its similarity to the user query. As mentioned above, more importance is assigned to the initial set of detected concepts than the additional concepts.

For each user query, the query processor can issue several visual queries to the typical CBIR system, one for the initial user query and one for each relevant concept. The feature centroids of concepts are used as their visual queries. The query processor then merges the results from the typical CBIR system to all the visual queries into a unique list for the user. A unique score is assigned to each result image using its rank(s) in the

result(s) and the score(s) of the relevant concept(s) from which the corresponding visual queries were generated. The final rank of each image is the minimum rank of the image across the results to the individual queries. The ranks of the results for an individual query are scaled by the score of the concept from which the query was generated.

7.5 Evaluation Experiments

In this section, we present the setup and discuss the results of the experiments performed for evaluating the MediaNet retrieval system in searching for images. In particular, we compared the performance of the MediaNet retrieval system with that of the typical CBIR system. Both systems are shown in Figure 7.1. For the MediaNet retrieval system, we distinguished two cases: visual and textual queries. The retrieval effectiveness was measured in terms of precision and recall for 51 queries using a collection of 5466 images.

7.5.1 Setup

The collection was 5466 images used in MPEG-7 to evaluate and compare color description technology [173]. 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. Some of the images included labels. Examples of images with their labels in the MPEG-7 color set are shown in Figure 7.4.

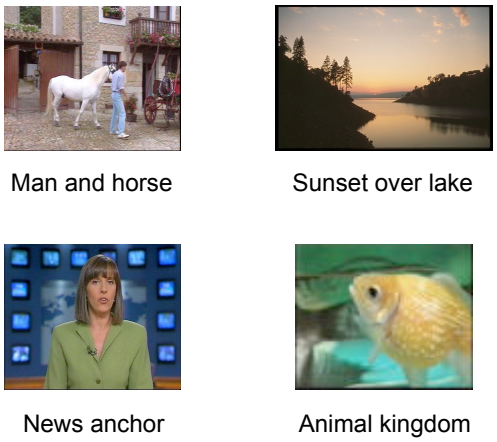


Figure 7.4: Examples of images with labels in the MPEG-7 color set.

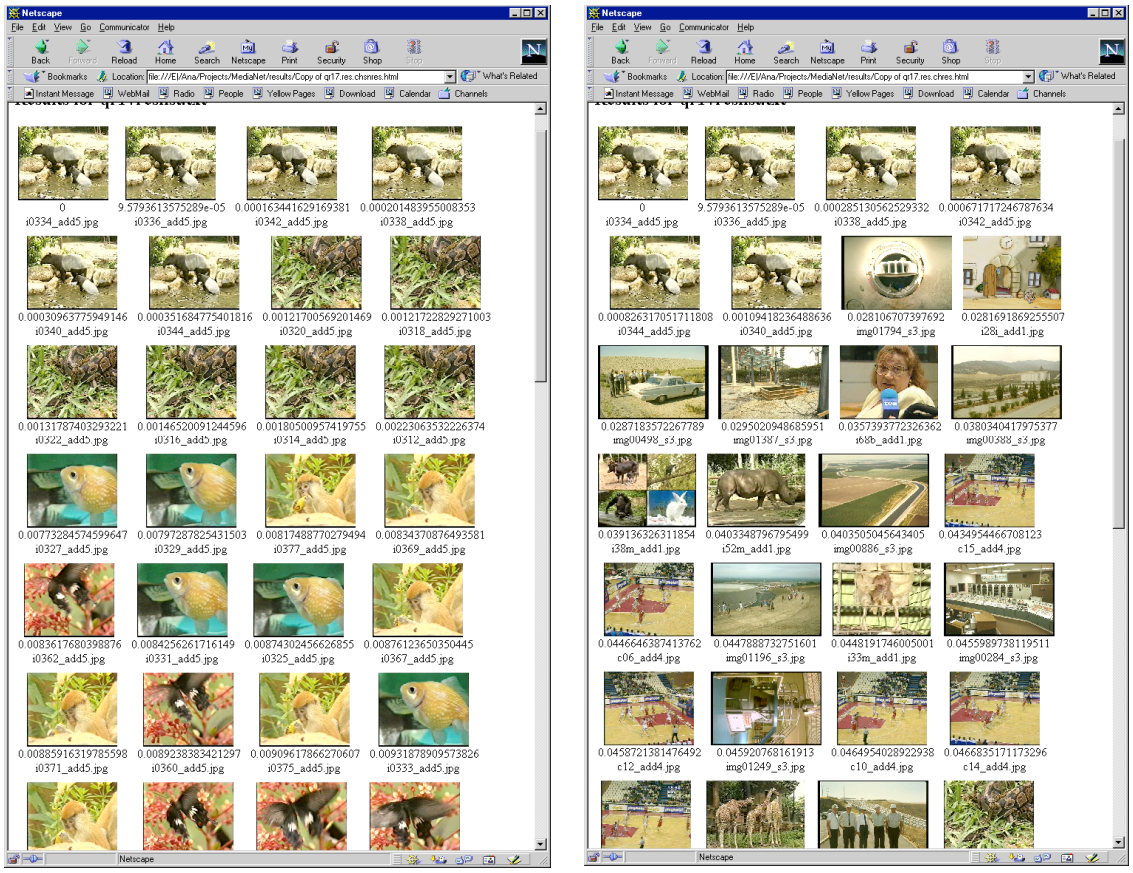
The queries and the ground truth for 50 queries were generated by MPEG-7 to evaluate and compare different color features proposed to the standard. Each query was annotated with short textual descriptions (e.g. "Flower Garden" and "News Anchor"). The focus of the ground truth was on color similarity so we also selected a semantic query and generated its ground truth manually. We picked the semantic query "Tapirs" as the keyword query "tapirs" or as the visual query of an image depicting a tapir. Relevance scores for calculating the retrieval effectiveness [132] were assigned to images in the collection for this query 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.

The medianet was constructed using the textual annotations of the 50 queries from MPEG-7 and half of the images in the ground truth, about 196 images. The rest of the images were used to build the feature database in the typical CBIR system. The total

number of concepts derived from the annotations was 96. 50 of these concepts were related to other concepts by specialization relationships; 34 concepts, by composition relationships. There was only one case of opposite relationship. The images used to construct the medianet were used to generate the feature centroids of the concepts in the medianet. Two different sets of visual features were used in the construction of the medianets and in the feature database in the typical CBIR system. The two sets were (1) the color histogram; and (2) color histogram, color coherence, wavelet texture, and Tamura texture.

7.5.2 Results

Figure 7.6 shows the average precision and recall for the MediaNet retrieval system and the typical CBIR system for the 50 MPEG-7 queries and the semantic query "Tapirs" for the first feature set, color histogram. The results using the second set of features, color histogram, color coherence, wavelet texture, and Tamura texture, were very similar. Recall and precision are standard measures used to evaluate the effectiveness of a retrieval system [71]. Recall is defined as the percentage of relevant images that are retrieved. Precision is defined as the percentage of retrieved images that are relevant. Figure 7.5.a and Figure 7.5.b depict the first 28 images returned by the MediaNet retrieval system and the typical CBIR system, respectively, to an image query depicting a tapir.



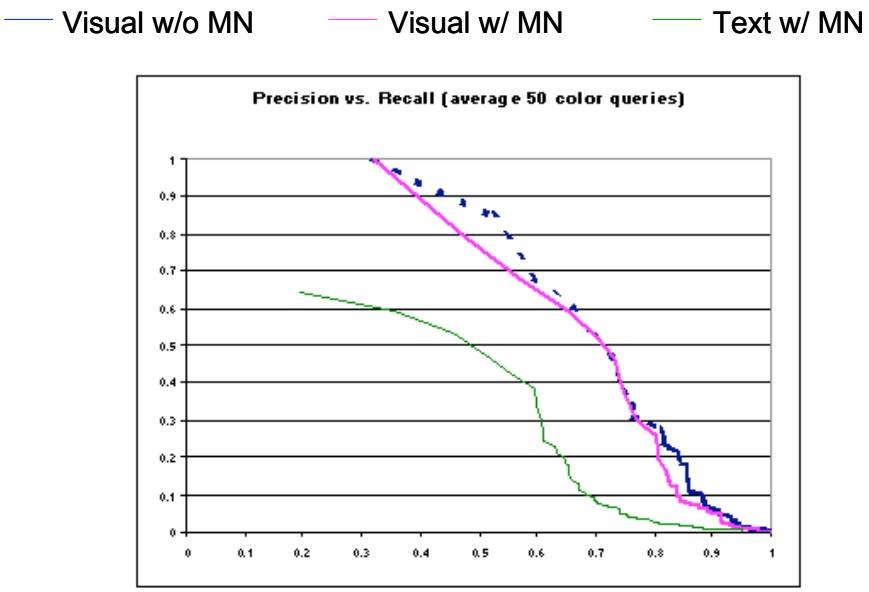
a) b)

Figure 7.5: The first 28 images returned by a) the MediaNet retrieval system and b) the typical CBIR system to an image query depicting a tapir.

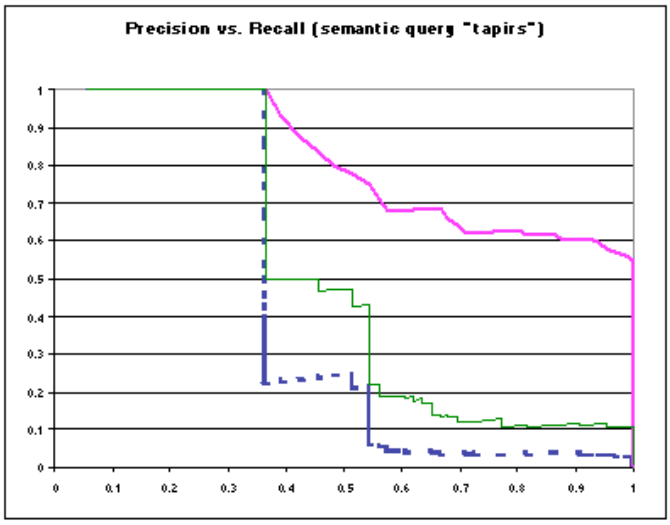
7.5.3 Discussion

As shown in Figure 7.6, for image queries, the performance of the MediaNet retrieval system and the typical CBIR system is comparable for the 50 MPEG-7 queries; however, the MediaNet retrieval system has superior retrieval effectiveness for the semantic query "Tapirs". Figure 7.5 demonstrates the different nature of both retrieval systems. Whereas the typical CBIR system returns image visually similar to the query image (see Figure

7.5.b); the MediaNet retrieval system returns images that are both semantically related and visually similar to input queries (see Figure 7.5.a). The retrieval effectiveness for textual queries in the MediaNet retrieval system is lower than for image queries due to the small number of concepts in the medianet. These results are very encouraging; however, additional experiments are needed to demonstrate the performance gain of using the MediaNet retrieval system.



a)



b)

Figure 7.6: Average precision and recall for a) the 50 MPEG-7 queries and b) the semantic query "Tapirs" using color histogram. "Visual w/o MN" corresponds to the typical CBIR system that does not use any medianet or query processor; "Visual w/ MN" to the MediaNet retrieval system with image queries; "Text w/ MN" to the MediaNet retrieval system with textual queries.

7.6 Summary

We have presented innovative ways of retrieving images based on medianets extracted from annotated images.

We have proposed to extend typical Content-Based Image Retrieval (CBIR) systems with medianets and advanced query processors. The medianet, which can include both semantic and perceptual concepts, is separated from the collection from which images are retrieved based on extracted features. The query processor translates, expands, and/or refines incoming queries across different media, as needed, using the medianet.

The transformed queries are then inputted to a typical CBIR system and the results merged by the query processor.

Initial experiments have demonstrated the superior retrieval effectiveness of the MediaNet retrieval system compared to a typical CBIR system. In particular, we observed the values of both recall and precision to be much higher for the semantic query "Tapirs" using the MediaNet retrieval system. The ground truth for this query was generated assigning non-null scores to images with mammals and other animals. The performance of the two systems was very similar for 50 queries created by MPEG-7 to evaluate and compare color features proposed to MPEG-7. Although these results are encouraging, additional experiments are needed to demonstrate the performance gain of the MediaNet retrieval system compared to typical CBIR systems.

Another interesting issue for future work is to update dynamically the weights of the relationships for calculating concept distances. The weights could be adapted and personalized based on user feedback and concept distances obtained as described in section 4.5.1. In the future, we also plan to investigate more advanced methods for detecting relevant concepts in images and merging the results in the query processor. For example, the image classification techniques presented in chapter 5 could be used for these purposes. Other future work will be to explore issues involved in querying large, distributed CBIR systems (image meta-search engines) using knowledge extracted from annotated images and WordNet, among others. Some of our relevant prior work on meta-search engines for images is presented in [8][9][30]. We have already implemented

such a system in the IMKA framework but it has not been evaluated yet (see section 8.3.2).

8 The IMKA Framework

8.1 Introduction

This chapter presents a framework for designing, configuring, testing, and evaluating generic multimedia information systems, the IMKA framework. This framework implements and uses the methods presented in the previous chapters of this thesis.

In recent years, there has been a major increase in available multimedia and in technologies to access multimedia. Users need multimedia information systems that enable the effective and efficient organization, retrieval, filtering, and browsing of multimedia satisfying users' needs. However, most current multimedia information systems are computationally expensive and have rigid infrastructures. Replacing components (e.g., the interface) and modifying the configuration (e.g., adding new feature databases) to these systems is usually an arduous task that involves careful programming.

This chapter focuses on a flexible framework for designing, configuring, testing, and evaluating generic multimedia information systems. There is very little work about this topic in the multimedia literature. One of the objectives of the MPEG-7 standard for multimedia description was to improve the flexibility, modularity, and scalability of multimedia applications. However, the actual implementation of the standard by the

MPEG-7 community, known as the Experimental Model [61], did not particularly have these qualities: there was little reusability of components and their functionality, in many instances, was restricted to special cases.

In this chapter, we present a novel and flexible system for developing and evaluating multimedia information systems based on medianets, the IMKA framework [18]. The IMKA framework supports several components that can be connected and configured to build different systems. The components currently supported are feature databases, search engines, meta-search engines, and medianet databases, among others. Figure 8.1 depicts a simple system in the IMKA framework with one search engine and two feature databases, a color histogram database and a Tamura texture database. The search engine retrieves images based on a combination of color histogram and Tamura texture features for the user. Both the MediaNet browsing and retrieval systems described in chapters 6 and 7, respectively, were built in the IMKA framework. The IMKA framework is implemented using the programming language Java. Users can currently build systems by instantiating, configuring, and connecting the classes corresponding to the system components in a Java program.

8.1.1 Outline of the Chapter

The rest of the chapter is organized as follows. In section 8.2, we describe the IMKA framework, in particular, the functional components and the software architecture. We

present examples of multimedia information systems built using the IMKA framework in section 8.3. Finally, we conclude with a summary of the chapter in section 8.4.

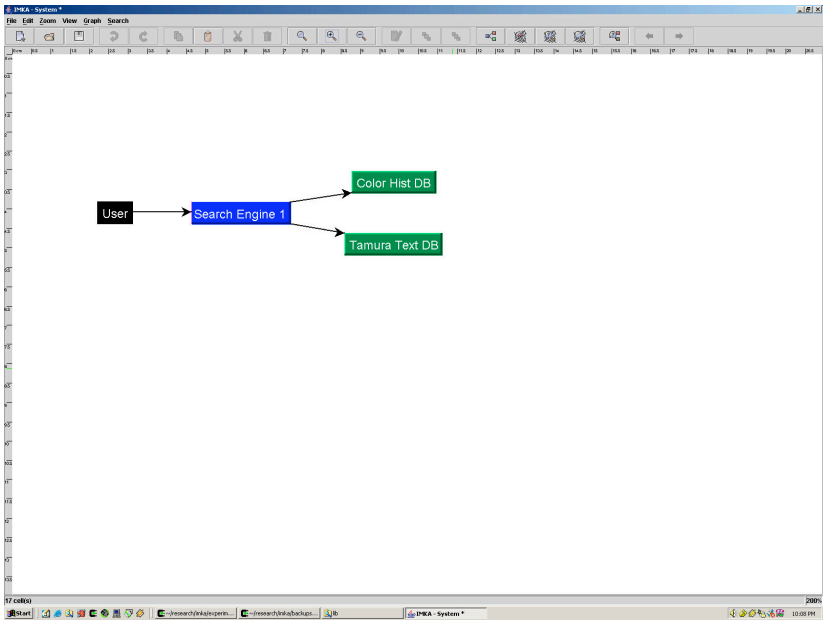


Figure 8.1: Simple example of an IMKA system with the user ("User"), one search engine ("Search Engine 1"), and two feature databases, a color histogram database ("Color Hist. DB") and tamura texture database ("Tamura Text. DB").

8.2 The IMKA Framework

This section describes the main functional components and the software architecture of the IMKA framework. The IMKA framework is a flexible framework for designing, configuring, testing, and evaluating generic multimedia information systems. In this framework, diverse multimedia information systems can be built by interconnecting system components (e.g., feature databases and search engines) in different configurations using the Java programming language.

8.2.1 Functional Components

The IMKA framework supports typical components for building Content-Based Image Retrieval (CBIR) systems, i.e., feature databases and search engines. A feature database contains a list of feature vectors extracted from images or their annotations. There is one feature vector per image in a feature database; the feature vectors for all the images are of the same type. For example, a color histogram database has a color histogram vector for every image in the collection. A feature database offers the functionality of extracting the specific feature from an image or its annotations, and returning the most similar images in response to a query (e.g., an image or a feature vector of the same type). A search engine can be connected to several feature databases. It forwards incoming queries to the feature databases and merges the results from each feature database. The queries are represented using medianets; they can have images, text, and feature vectors, among others. System components process incoming queries to filter out the irrelevant components they cannot match. For example, a color histogram database only keeps and matches images or color histogram feature vectors from incoming queries.

Additional system components supported by the IMKA framework are meta-search engines, medianet databases, and query processors. These specialized components support, among others, the construction of the novel MediaNet browsing and retrieval systems described in chapters 6 and 7, respectively (see section 8.3). A meta-search engine can be connected to several search engines. It forwards incoming queries to, and merges the results from, each connected search engine. A medianet database contains a

concept network with media examples (i.e., a medianet) constructed for a collection of annotated images as described in chapter 4. Query processors can translate, expand, and/or refine queries across different media using a medianet, as described in chapter 7. Query processors can submit the processed queries to search engines and feature databases, and merge the results from those components. Therefore, apart from medianets, query processors can be connected to meta-search engines, search engines, and feature databases, among others. A special kind of medianet database, the hierarchical medianet database, contains a hierarchy of medianets that can browsed together with the image collection as described in chapter 6.

Queries can be submitted to any IMKA system. The query and the results evolve from one component to the next in the system. That is why the IMKA framework enables to view the actual query and the results at each component. In addition, the knowledge of each component can be viewed and browsed in the form of a medianet. The knowledge of a color histogram database is a list of images with the corresponding color histogram feature vectors in the database. The knowledge of a hierarchical medianet database is a hierarchy of medianets that can be browsed using the MediaNet browser (see chapter 6). Several examples of IMKA systems are presented in the section 8.3.

8.2.2 Software Architecture

The IMKA framework is implemented using the programming language Java. Each system component described above is defined as a class forming a hierarchy derived

from the abstract class SystemComponent. Figure 8.2 shows the class hierarchy of all system components currently available in the IMKA framework.

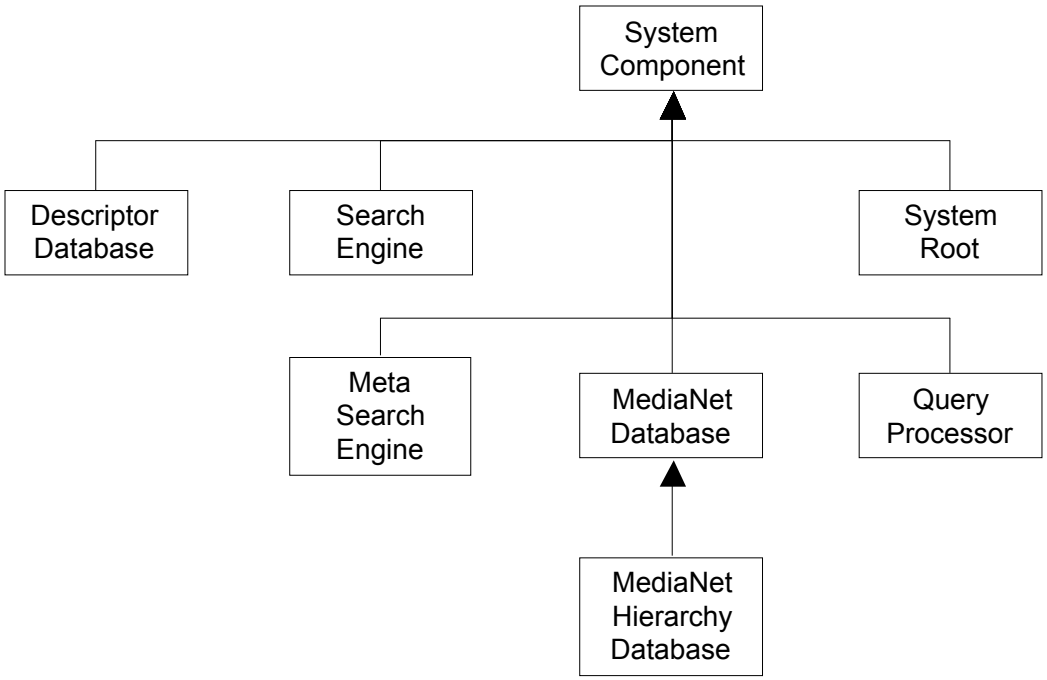


Figure 8.2: Class hierarchy of the system components in the IMKA framework. UML conventions are used to represent the relationship between classes: the arrow represents the inheritance relationship (e.g., the DescriptorDatabase class extends SystemComponent class in Java terminology).

The functionality of each class in Figure 8.2 follows the descriptions in section 8.2.1. In addition, the SystemRoot class represents the user: the recipient of all the functionality and knowledge accumulated by the systems components connected to it. For example, see the User component in Figure 8.3 and Figure 8.4, which is connected to a MediaNet hierarchy database (with name "Hierarchical Concept Network DB") and a query processor (with name "Query Processor"), respectively.

An IMKA system is built by instantiating, configuring, and connecting the classes of the corresponding and desired the system components in a Java program. Examples are provided in the next section. IMKA systems can be saved/read to/from files by the IMKA application, which displays them graphically as shown in Figure 8.3 and Figure 8.4. In the future, the IMKA application will be extended so that users can build systems by dragging, dropping, configuring, and connecting the systems components directly in the graphical user interface.

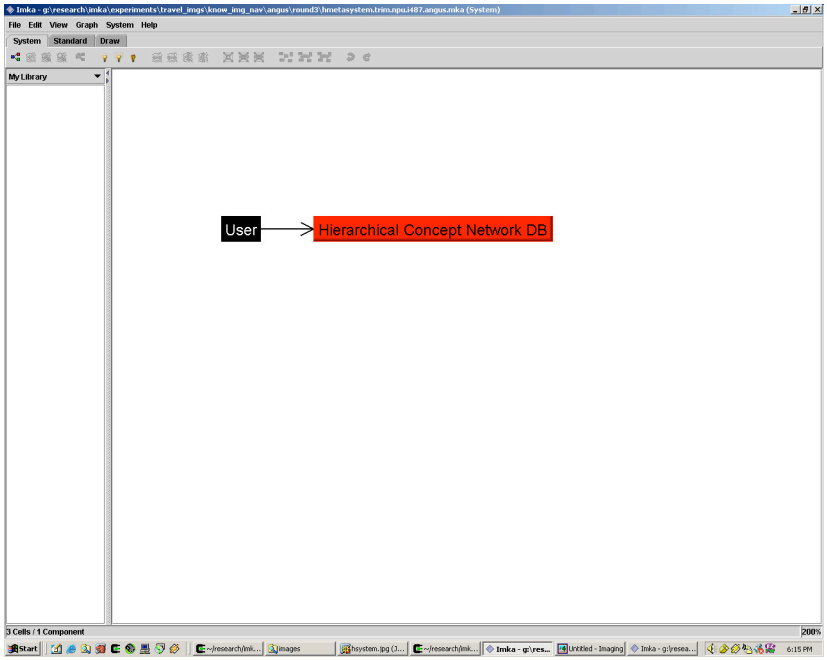


Figure 8.3: Components and configuration of the MediaNet browser in the IMKA framework. It is composed of the user ("User") and a medianet hierarchy database ("Hierarchical Concept Network DB").

8.3 Examples of IMKA Systems

Both the MediaNet browsing and retrieval systems described in chapters 6 and 7, respectively, were built in the IMKA framework. Figure 8.3 and Figure 8.4 show the components and the configuration of each system, which are further described in this section.

8.3.1 MediaNet Browsing System

The MediaNet browsing system proposed in chapter 6 was built in the IMKA framework. The components and the configuration of the system are shown in Figure 8.3.

As shown in Figure 8.3, this is a very simple system composed of a hierarchical medianet database connected directly to the user. The hierarchy of medianets was built for a collection of annotated images as described in section 6.3. No queries can be served by this system. However, the system enables users to browse the image collection and the medianet hierarchy using the advanced visualization and navigation techniques presented in chapter 6. This was the system with which the subjects of the user study interacted during the evaluation of the MediaNet browser. The WordNet browser, which was compared to the MediaNet browser, was implemented as a special hierarchical medianet database that only contained a concept hierarchy. An abstract of the code in Java that instantiates, configures, and connects the components for the IMKA system in Figure 8.3 is included in Table 8.1.

Table 8.1: Abstract of code in Java that instantiates, configures, and connects the components for the IMKA system in Figure 8.3.

```
// IMKA System
MetaSystem system = new MetaSystem();
Set insert = new HashSet();
ConnectionSet connections = new ConnectionSet();

// Create and configure User component
SystemRoot root = new SystemRoot();
root.setString("User");
insert.add(root);

// Create and configure Concept Network Hierarchical Database component
MediaNetHierarchyDatabase mnpdb = new MediaNetHierarchyDatabase();
mnpdb.setMediaNet(aMediaNet);
mnpdb.setString("Hierarchical Concept Network DB");
insert.add(mnpdb);

// Edge connecting system components
DefaultEdge edge = new DefaultEdge("Edge");
insert.add(edge);
connections.connect(edge, root.getDefaultPort(), mnpdb.getDefaultPort());

// Add components and connections to the IMKA system
system.insert(insert.toArray(), null, connections, null, null);
```

8.3.2 MediaNet Retrieval System

The MediaNet retrieval system proposed in chapter 7 with an additional meta-search engine was also built in the IMKA framework. The components and the configuration of the system are shown in Figure 8.4.

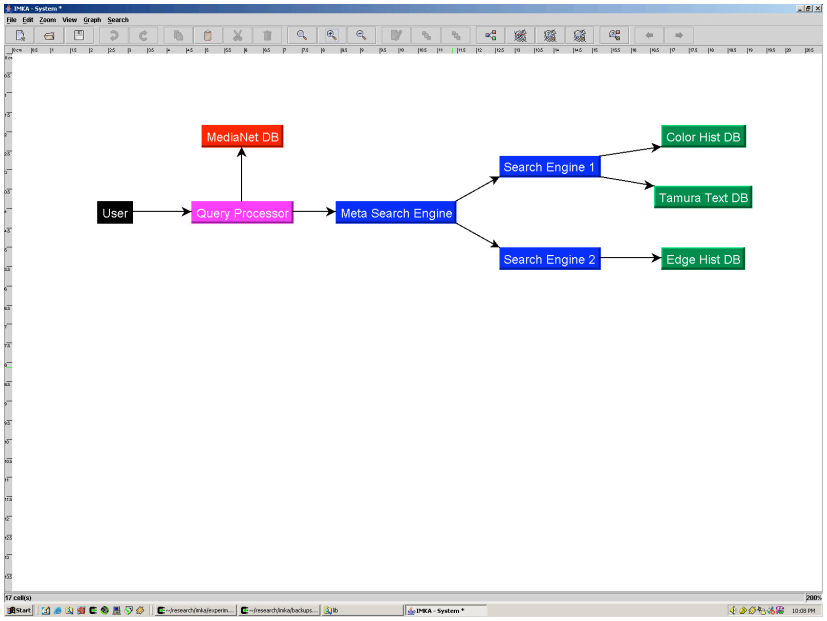
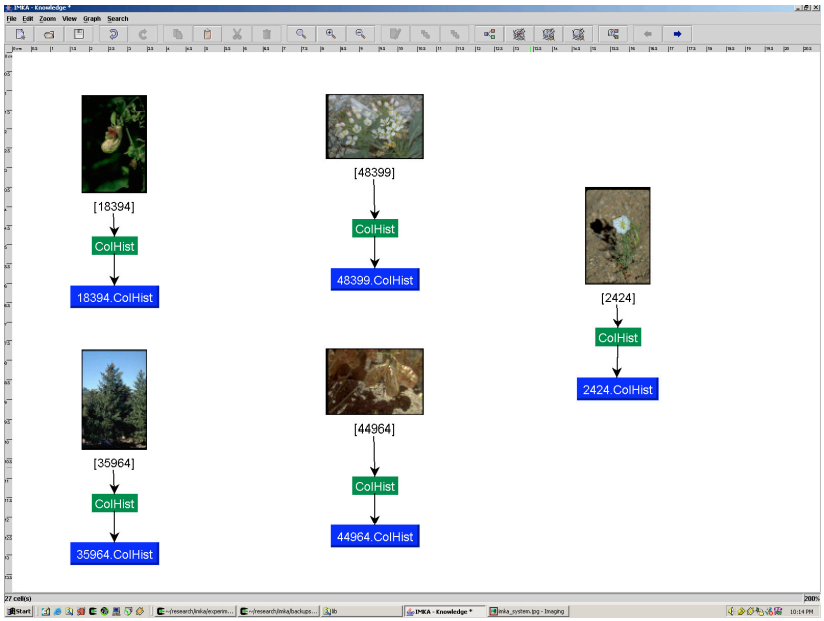


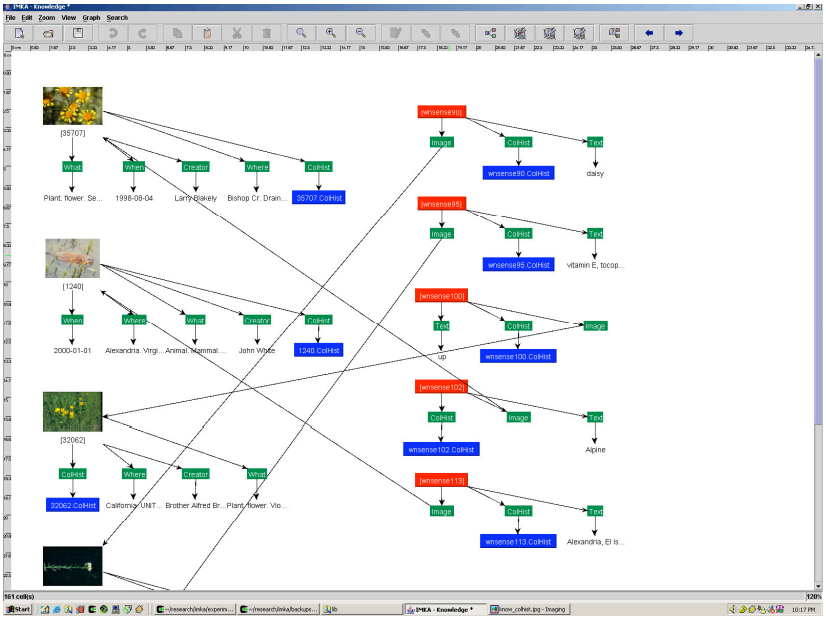
Figure 8.4: Components and configuration of the MediaNet retrieval system in the IMKA framework. It is composed of the user ("User"), a query processor ("Query Processor"), a medianet ("MediaNet DB"), a meta-search engine ("Meta Search Engine"), two search engines ("Search Engine 1" and "Search Engine 2"), and three feature databases for color histogram, Tamura texture, and edge direction histogram ("Color Hist. DB", "Tamura Text. DB", and "Edge Hist.DB", respectively).

As shown in Figure 8.4, the MediaNet retrieval system is composed of three feature databases, two search engines, one meta-search engine, a medianet database, and a query processor connected to the user. The query processor translates, extends, and/or refines

user queries across different media using a medianet, and submits them to the meta-search engine. In a similar way, the processed query is forwarded from the meta-search engine to two search engines, and from the search engines to the three feature databases. The results from the feature databases are merged and integrated starting at the search engines to the query processor through the meta-search engine. Queries can be submitted to the entire system or to specific components in the system. As mentioned before, the query, the results, and the knowledge of each component can be visualized using the IMKA framework. Figure 8.5.a and Figure 8.5.b show the knowledge at the color histogram database and at the medianet database in the IMKA system shown in Figure 8.4. Figure 8.6 shows the query at the query processor for a user query. Figure 8.7 shows the results at the color histogram database. An abstract of the code in Java that instantiates, configures, and connects the components for the IMKA system in Figure 8.4 is included in Table 8.2.



a)



b)

Figure 8.5: Knowledge a) at the color histogram database and b) at the medianet database in the IMKA system shown in Figure 8.4.

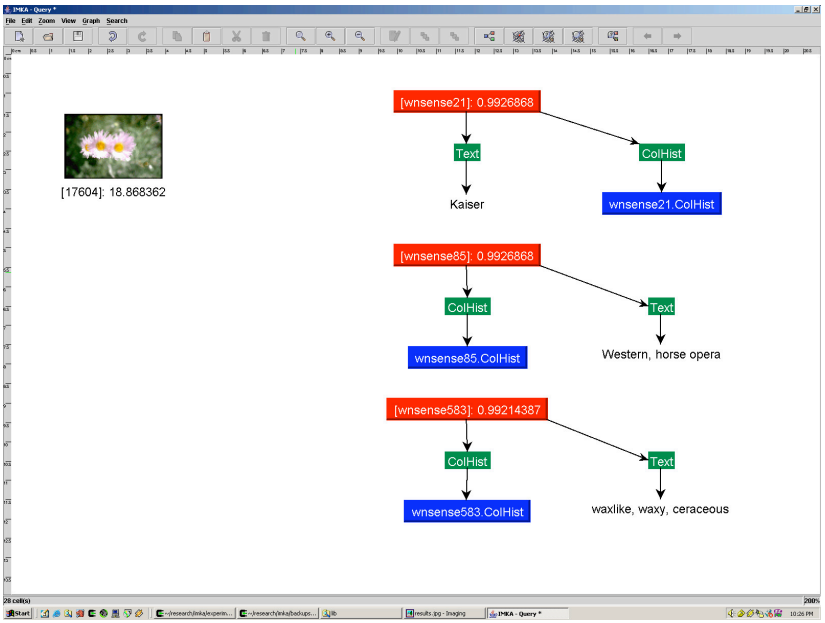


Figure 8.6: Queries at the query processor in the IMKA system shown in Figure 8.4. The user query was the image depicting the flower.

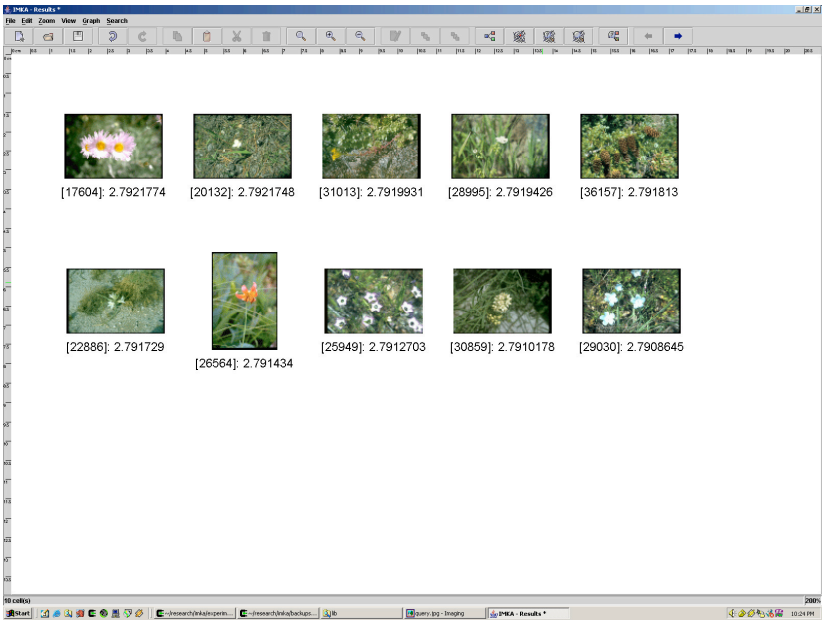


Figure 8.7: Results of the query in Figure 8.6 at the user component in the IMKA system shown in Figure 8.4.

Table 8.2: Abstract of code in Java that instantiates, configures, and connects the components for the IMKA system in Figure 8.4.

```
// IMKA System
MetaSystem system = new MetaSystem();
Set insert = new HashSet();
ConnectionSet connections = new ConnectionSet();

// Create and configure User component
SystemRoot root = new SystemRoot();
root.setString("User");
insert.add(root);

// Create and configure Query Processor component
QueryProcessor qp1 = new QueryProcessor();
qp1.setString("Query Processor");
insert.add(qp1);

// Create and configure MediaNet Database component
MediaNetDatabase mndb1 = new MediaNetDatabase();
mndb1.setImgList(anImageList);
mndb1.addDescDatabase(aColorHistDB);
mndb1.addDescDatabase(aTamuraTextureDB);
mndb1.addClusterDatabase(aClusteringDB);
mndb1.addAntDatabase(anAnnotationDB);
mndb1.addWNWordDatabase(aDisambiguatedWordDB);
mndb1.setString("MediaNet DB");
insert.add(mndb1);
```

```
// Create and configure Meta Search Engine component
MetaSearchEngine mse1 = new MetaSearchEngine();
mse1.setString("Meta Search Engine");
insert.add(mse1);

// Create and configure Search Engine components
SearchEngine se1 = new SearchEngine();
se1.setString("Search Engine 1");
insert.add(se1);

SearchEngine se2 = new SearchEngine();
se2.setString("Search Engine 2");
insert.add(se2);

// Create and configure Descriptor Database components
DescriptorDatabase db1 = new DescriptorDatabase(
      "Color Hist", aColorHistDB, anImageList);
db1.setString("Color Hist DB");
insert.add(db1);

DescriptorDatabase db2 = new DescriptorDatabase(
      "Tamura Text ", aTamuraTextureDB, anImageList);
db2.setString("Tamura Text DB");
insert.add(db2);

DescriptorDatabase db3 = new DescriptorDatabase(
      "Edge Hist", anEdgeHistDB, anImageList);
db3.setString("Edge Hist DB");
insert.add(db3);
```

```
// Edges connecting system components
DefaultEdge e = new DefaultEdge("Edge1");
insert.add(e);
connections.connect(e, root.getDefaultPort(), qp1.getDefaultPort());

e = new DefaultEdge("Edge2");
insert.add(e);
connections.connect(e, qp1.getDefaultPort(), mndb1.getDefaultPort())

e = new DefaultEdge("Edge3");
insert.add(e);
connections.connect(e, qp1.getDefaultPort(), mse1.getDefaultPort());

e = new DefaultEdge("Edge4");
insert.add(e);
connections.connect(e, mse1.getDefaultPort(), se1.getDefaultPort());

e = new DefaultEdge("Edge5");
insert.add(e);
connections.connect(e, mse1.getDefaultPort(), se2.getDefaultPort());

e = new DefaultEdge("Edge6");
insert.add(e);
connections.connect(e, se1.getDefaultPort(), db1.getDefaultPort());

e = new DefaultEdge("Edge7");
insert.add(e);
connections.connect(e, se1.getDefaultPort(), db2.getDefaultPort());
```

```
e = new DefaultEdge("Edge8");
insert.add(e);
connections.connect(e, se2.getDefaultPort(), db3.getDefaultPort());

// Add components and connections to the IMKA system
system.insert(insert.toArray(), null, connections, null, null);
```

8.4 Summary

We have presented the IMKA framework, a novel and flexible framework for developing and evaluating multimedia information systems based on medianets. The IMKA framework supports several components that can be connected and configured to build complex multimedia information systems. The components currently supported are feature databases, search engines, meta-search engines, medianet databases, and query processors, among others. The feasibility of this framework was demonstrated by building two real multimedia information systems, the MediaNet browser and retrieval system described in chapters 6 and 7, respectively. The IMKA framework is implemented in a modular way using the programming language Java.

9 Conclusions and Future Work

9.1 Introduction

In this chapter, we summarize work presented in this thesis and discuss directions for future work.

We have addressed several challenging problems in representing and discovering multimedia knowledge, and in classifying, browsing and retrieving multimedia using the extracted knowledge. This thesis has proposed solutions to these problems that invariably integrate perception and semantics. For example, we have integrated the processing of images and words, and combined perceptual and semantic knowledge in the same framework, among others. The intuition behind these approaches has come from human mental models, which seem to contain nodes of not only textual (semantic) but also audio-visual nature (perceptual).

9.1.1 Outline of the Chapter

In section 9.2, we briefly summarize the work done in this thesis. Finally, we conclude with some interesting directions for future work in section 9.3.

9.2 Summary of the Thesis

This section briefly summarizes the work presented in this thesis on multimedia knowledge representation and discovery, and knowledge-based image classification, browsing and retrieval. All the techniques and methods proposed in this thesis have been implemented in the IMKA framework, which also serves as a framework for designing, configuring, testing, and evaluating generic multimedia information systems.

9.2.1 Multimedia Knowledge Representation

This thesis has developed a novel framework for representing generic perceptual and semantic knowledge using multimedia, MediaNet (see chapter 3). The main components of the MediaNet framework include concepts, relationship between concepts, and media exemplifying concepts and relationships such as images, words, and low-level features of the media.

In designing the MediaNet framework, we have built on the basic principles of semiotics and semantic networks, in addition to utilizing evidence from psychology that humans have rich mental models of the world that contain interconnected nodes of textual and audio-visual nature. We have also proposed to use MPEG-7 structured collection, structure, and semantic description tools to encode and represent medianets in a machine processable, reusable, and interoperable form. The concept network is described using the structure collection description tools; whereas, structure and semantic description tools are used to represent the media examples in the medianet.

9.2.2 Multimedia Knowledge Discovery

This thesis has proposed novel methods for automatically discovering, summarizing, and evaluating multimedia knowledge from annotated images in the form of image clusters, word senses, and relationships, among them (see chapter 4). These are essential for applications to intelligently, efficiently, and coherently deal with multimedia.

The proposed methods include (1) new techniques for discovering statistical and similarity relationships among image clusters (perceptual relationships and concepts); (2) a novel technique for disambiguating the senses of words and their relationships in image annotations that uses not only the annotations and WordNet but also the image clusters (semantic concepts and relationships); (3) a new technique for calculating distances between concepts used to cluster concepts for summarizing medianets; and (4) automatic ways of measuring the consistency, completeness, and conciseness of medianets using notions from information and graph theory.

Experiments have shown both visual and textual features are useful in extracting different perceptual knowledge from annotated images; therefore, the integration of both kinds of features has potential to improve performance compared to individual features. The image clusters based on both visual and textual features have been shown to have a higher correlation with some of the considered categories than image clusters based on either visual or textual features (about 8% gain for 64 clusters). The evaluation of the proposed word-sense disambiguation approach has shown that using perceptual

knowledge in the form of image clusters can improve performance compared to most frequent senses and text-based word-sense disambiguation (about 6% gain for nature images). Additional experiments have shown the importance of good concept distance measures, as the proposed one, for clustering and summarizing medianets, and the potential of the proposed automatic measures for measuring the quality of medianets.

9.2.3 Knowledge-Based Image Classification

This thesis has presented novel methods for classifying images based on medianets discovered from annotated images (see chapter 5). The novelty of this work is the automatic class discovery and the classifier combination using extracted medianets.

The extracted medianets are networks of concepts (e.g., image clusters and word-senses) with associated text and images, which are built automatically from annotated images using the electronic dictionary WordNet. Concepts that are similar statistically are merged to reduce the size of the medianet. Our MediaNet classifier is constructed by training a meta-classifier to predict the presence of each concept in images. A Bayesian network is then learned using the meta-classifiers and the concept network. For a new image, the presence of concepts is first detected using the meta-classifiers and refined using Bayesian inference.

Experiments have shown that classifiers based on medianets discovered and summarized from annotated images using human knowledge (i.e., WordNet) can result in superior accuracy to individual classifiers and purely statistically learned classifier structures. The

improvement in accuracy for classifiers that are not originally very good is significant (up to 15-30% for NB classifiers); however, the gains for originally accurate classifiers are more modest (up to 4% for SVM classifiers). Another contribution of this work is the analysis of the role of visual and textual features in image classification. As textual or joint visual-textual features perform better in classifying images than visual features, we try to predict textual features for images without annotations; however, we have found that the accuracy of visual-predicted textual features does not consistently improve over using only visual features.

9.2.4 Knowledge-Based Image Browsing

This thesis has also presented novel techniques for automatically organizing and browsing annotated images based on extracted and summarized knowledge in the form of hierarchies of medianets (see chapter 6).

Annotated images are organized in medianet hierarchies (i.e., concepts network hierarchies), which are constructed based on knowledge extracted and summarized from images and annotations. The initial medianet discovered from the collection is clustered hierarchically resulting in a medianet hierarchy. The discovered medianet includes semantic knowledge (e.g., image clusters and relationships), semantic knowledge (e.g., word sense and relationships), and statistical interrelations among these. Users can then browse the annotated images by navigating the resulting medianet hierarchy. Ideas from fish-eye views and spring modeling are exploited for displaying concepts using text and

images, and for drawing networks, respectively. In spite of the medianet hierarchy, its navigation does not necessarily need to be strictly hierarchical.

An extensive task-oriented, user-centered study has demonstrated the superiority of the proposed techniques in terms of subjective satisfaction (e.g., more useful, easier, more stimulating, and more successful), efficiency (e.g., fewer viewed images and executed browsing operations), and effectiveness (e.g., better images selected for two travel pamphlets) of users in performing common browsing tasks such as searching for images related to a specific topic. We have compared the performance our approach with the thesaurus-based image browsing system proposed by Yang et al. [165] through the administration of questionnaires to subjects and the monitoring of system parameters in the background (e.g. execution time), respectively.

9.2.5 Knowledge-Based Image Retrieval

We have also proposed in this thesis innovative ways of retrieving images based on medianets extracted from annotated images (see chapter 7). We have proposed to extend typical Content-Based Image Retrieval (CBIR) systems with medianets and advanced query processors. The medianet, which can include both semantic and perceptual concepts, is separated from the collection from which images are retrieved based on extracted features. The query processor translates, expands, and/or refines incoming queries across different media, as needed, using the medianet. The transformed queries

are then inputted to a typical CBIR system and the results merged by the query processor.

Initial experiments have demonstrated the superior retrieval effectiveness of the MediaNet retrieval system compared to a typical CBIR system. In particular, we observed the values of both recall and precision to be much higher for the semantic query "Tapirs" using the MediaNet retrieval system. The ground truth for this query was generated assigning non-null scores to images with mammals and other animals. The performance of the two systems was very similar for 50 queries created by MPEG-7 to evaluate and compare color features proposed to MPEG-7. Although these results are encouraging, additional experiments are needed to demonstrate the performance gain of the MediaNet retrieval system compared to typical CBIR systems.

9.3 Future Work

The problems addressed by this thesis are very challenging. This thesis aims at providing an innovative and complete solution to building multimedia information that better satisfy the users' demands and needs. In this section, we discuss some of the remaining issues in our proposed solutions.

9.3.1 Multimedia Knowledge Representation

There is a clear trend of extending the current Web so that information is given well-defined meaning, better enabling the automation of many services, and the cooperation

of computers and people. The approach taken by W3C is to promote the use of semantic markup languages such as RDF (Resource Description Framework) and OWL (Web Ontology Language) for publishing and sharing information and ontologies on the Web. As future work, we plan to define the encoding and representation of medianets using these markup languages. In addition, we plan to generate medianets for different ontologies and make them online so that

9.3.2 Multimedia Knowledge Discovery

Large amounts of current multimedia are in the form of audio-visual data with optional caption and subtitle channels. For example, CNN receives over 300 hours of raw footage every day. In addition, there are many external resources of useful knowledge that are underutilized in multimedia information systems such as the Internet itself and the Cyc ontology. In the future, we plan to expand the multimedia knowledge discovery work in two directions: a) developing algorithms for other media such as moving pictures and audio, and b) integrating knowledge from other external resources apart from WordNet. As most multimedia collections are dynamic and grow with time, future work will also consist of proposing methods for modifying and updating the extracted medianets when images are added or removed from the collection. In addition, new procedures are needed to evaluate the quality of the knowledge represented by medianets, i.e., specific instances of the MediaNet framework.

9.3.3 Knowledge-Based Image Classification

Many times the labels or the words associated with images only apply to specific regions of the images. Imagine the picture of a tiger in the wild; the label "tiger" is only applicable to the region that depicts the tiger. In the future, we plan to develop algorithms for distinguishing between concepts that are applicable to entire images (e.g., "outdoor") or only to regions within images (e.g., "tiger"). We are also interested in the problem of classifying image regions using annotations assigned globally to images. Some initial work in this area can be found in the literature [5][47][106].

9.3.4 Knowledge-Based Image Browsing

The user study in which we evaluated our knowledge-based approach to image browsing, i.e., the MediaNet browser, pointed out the subjectivity (1) in the annotation of images and (2) in the appreciation of image similarity. As future work, we are interested in extending the MediaNet browser with controls and relevance feedback mechanisms so that the medianet hierarchy and its visualization can be personalized and adapted to individual users' preferences. The interface of the MediaNet browser is flexible enough to capture sophisticated feedback from users to modify, personalize, and optimize the medianet hierarchies (e.g., drag images and concepts around, and delete, paste, and copy nodes from one screen to another). The sophisticated user feedback could be used to modify, personalize, and optimize the medianet hierarchies in a much more powerful way than related systems that only allow users to modify clusters [33][62][7][81] or layouts

[125] of images. Another interesting direction for future work is to extend the user-centered, task-oriented design and evaluation of to other multimedia information systems.

9.3.5 Knowledge-Based Image Retrieval

Another interesting issue for future work is to update dynamically the weights of the relationships for calculating concept distances. The weights could be adapted and personalized based on user feedback and concept distances obtained as described in section 4.5.1. In the future, we also plan to investigate more advanced methods for detecting relevant concepts in images and merging the results in the query processor. For example, the image classification techniques presented in chapter 5 could be used for these purposes. Other future work will be to explore issues involved in querying large, distributed CBIR systems (image meta-search engines) using knowledge extracted from annotated images and WordNet, among others. Some of our relevant prior work on meta-search engines for images is presented in [8][9][30]. We have already implemented such a system in the IMKA framework but it has not been evaluated yet.

10 References

[1] Armitage, L., and P.G.B. Enser, "Analysis of User Need in Image Archives", Journal of Information Science, Vol. 23, No. 4, pp. 287-299, 1997.

[2] Aslandogan, Y.A., C. Their, C.T. Yu, and N. Rishe, "Using Semantic Contents and WordNet in Image Retrieval", in Proc. of Proc. of ACM Special Interest Group on Information Retrieval Conf., Conf. on Research and Development in Information Retrieval (ACM SIGIR-1997), pp. 286-295, Philadelphia, PA, USA, 1997.

[3] Bach, J.R., C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. Jain, and C.-F. Shu, "Virage Image Search Engine: An Open Framework for Image Management", in Proc. of IS&T/SPIE Conf. on Storage and Retrieval for Image and Video Databases IV, pp. 76-87, San Jose, CA, USA, 1996.

[4] Barnard, K., P. Duygulu, and D. Forsyth, "Clustering Art", in Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR-2001), Hawaii, USA, Dec. 9-14, 2001.

[5] Barnard, K., P. Duygulu, and D. Forsyth, N. de Freitas, D. Blei, and M.I. Jordan, "Matching Words and Pictures", Journal of Machine Learning Research (JMLR), Special Issue on Text and Images, Vol. 3, pp. 1107-1135, 2003.

[6] Barnard, K., M. Johnson, and D. Forsyth, "Word Sense Disambiguation with Pictures", in Proc. of Workshop on Learning Word Meaning from Non-Linguistic Data, held in conjunction with The Human Language Technology Conf., Edmonton, Canada, May 27-June 1, 2003.

[7] Bartolini, I., P. Ciaccia, and M. Patella, "The PIBE Personalizable Image Browsing Engine", in Proc. of Int. Workshop on Computer Vision meets Databases, (CVDB-2004), pp. 43-50, Paris, France, June 13, 2004.

[8] Beigi, M., A.B. Benitez, and S.-F. Chang, "MetaSEEk: A Content-Based Meta-Search Engine for Images", in Proc. of IS&T/SPIE Conf. on Storage and Retrieval for Image and Video Databases VI, Vol. 3312, San Jose, CA, USA, Jan 28-30, 1998.

[9] Benitez, A.B., M. Beigi, and S.-F. Chang, "Using Relevance Feedback in Content-Based Image Metasearch", IEEE Internet Computing, Vol. 2, No. 4, pp. 59-69, Jul/Aug 1998.

[10] Benitez, A.B., and S.-F. Chang, "Automatic Multimedia Knowledge Discovery, Summarization and Evaluation", submitted to IEEE Trans. on Multimedia, 2004;

also retrieved on Sept. 7, 2004, from http://www.ee.columbia.edu/dvmm/publications/03/IE3TMM_Ana03.pdf.

[11] Benitez, A.B., and S.-F. Chang, "Extraction, Description and Application of Multimedia Using MPEG-7", in Proc. Asimolar Conf. of Signals, Systems, and Computers, Invited Paper on Special Session on Document Image Processing, Monterey, CA, Nov. 2003.

[12] Benitez, A.B., and S.-F. Chang, "Image Classification Using Multimedia Knowledge Networks", in Proc. of IEEE Int. Conf. on Image Processing (ICIP-2003), Barcelona, Spain, Sep. 2003.

[13] Benitez, A.B., and S.-F. Chang, "Organization and Browsing of Annotated Images Using Multiresolution Knowledge Networks", ADVENT Technical Report #007 Columbia University, 2003.

[14] Benitez, A.B., and S.-F. Chang, "Multimedia Knowledge interrelation, Summarization and Evaluation", in Proc. Int. Workshop on Multimedia Data Mining, in conjunction with the Int. Conf. on Knowledge Discovery and Data Mining (MDM/KDD-2002), Edmonton, Alberta, Canada, July 23-26, 2002.

[15] Benitez, A.B., and S.-F. Chang, "Perceptual Knowledge Construction From Annotated Image Collections", in Proc. of Int. Conf. on Multimedia and Expo (ICME-2002), Lausanne, Switzerland, Aug 26-29, 2002.

[16] Benitez, A.B., and S.-F. Chang, "Semantic Knowledge Construction From Annotated Image Collections", in Proc. of Int. Conf. on Multimedia and Expo (ICME-2002), Lausanne, Switzerland, Aug 26-29, 2002.

[17] Benitez, A.B., and S.-F. Chang, "Validation Experiments on Structural, Conceptual, Collection, and Access Description Schemes for MPEG-7", in Digest of IEEE Int. Conf. on Consumer Electronics (ICCE-2000), Invited Paper on Special Session on MPEG-7, Los Angeles, CA, June 2000.

[18] Benitez, A.B., S.-F. Chang, and J.R. Smith, "IMKA: A Multimedia Organization System Combining Perceptual and Semantic Knowledge", in Proc. of ACM Int. Multimedia Conf. and Exhibition (ACM MM-2001), Ottawa, CA, Oct. 2001.

[19] Benitez, A.B., J.M. Martinez, H. Rising, P. Salembier, "Description of a Single Multimedia Document", In Introduction to MPEG 7: Multimedia Content Description Language, B. S. Manjunath, P. Salembier, T. Sikora (eds.), Chap. 8, pp. 111-138, Wiley, 2002.

[20] Benitez, A.B., S. Paek, S.-F. Chang, Q. Huang, A. Puri, C.-S. Li, J.R. Smith, L.D. Bergman, C.N. Judice, "Object-Based Multimedia Description Schemes and Applications for MPEG-7", Image Communication Journal (ICJ), Invited Paper on a Special Issue on MPEG-7, Vol. 16, No. 1, pp. 235-269, Sep. 2000.

[21] Benitez, A.B., H. Rising, C. Jorgensen, R. Leonardi, A. Bugatti, K. Hasida, R. Mehrotra, A.M. Tekalp, A. Ekin, T. Walker, "Semantics of Multimedia in MPEG-

7", in Proc. of IEEE Int. Conf. on Image Processing (ICIP-2002), Rochester, New York, Sep. 2002.

[22] Benitez, A.B., and J.R. Smith, "New Frontiers for Intelligent Content-Based Retrieval", in Proc. of the IS&T/SPIE Conf. on Storage and Retrieval for Media Databases, Vol. 4315, San Jose, CA, USA, Jan 24-26, 2001.

[23] Benitez, A.B., J.R. Smith, and S.-F. Chang, "MediaNet: A Multimedia Information Network for Knowledge Representation", in Proc. of IS&T/SPIE Conf. on Internet Multimedia Management Systems, Vol. 4210, Boston, MA, USA, Nov 6-8, 2000.

[24] Benitez, A.B., D. Zhong, S.-F. Chang, and J.R. Smith, "MPEG-7 MDS Content Description Tools and Applications", In Proc. of Int. Conf. on Computer Analysis of Images and Patterns (ICAIP-2001), Warsaw, Poland, Sep. 2001.

[25] Del Bimbo, A., "Expressive Semantics for Automatic Annotation and Retrieval of Video Streams", in Proc. of Int. Conf. on Multimedia and Expo (ICME-2000), New York, NY, July 2000.

[26] Bruning, R.H., G.J. Schraw, and R.R. Ronning, "Cognitive Psychology and Instruction", Prentice Hall, Englewood Cliffs, N.J., 1995.

[27] Budanitsky, A., and G. Hirst, "Semantic Distance in WordNet: An Experimental, Application-oriented Evaluation of Five Measures", in Proc. of North American

Association for Computational Linguistics (NAACL-2001), Pittsburgh, PA, June 2001.

[28] Campbell, N.W., and B.T. Thomas, "Automatic Segmentation and Classification of Outdoor Images Using Neural Networks", Int. Journal of Neural Systems, Vol. 8, No. 1, pp. 137-144.

[29] Chang, S.-F., W. Chen, H. J. Meng, H. Sundaram, D. Zhong, "A Fully Automated Content-Based Video Search Engine Supporting Spatio-Temporal Queries", IEEE Trans. on Circuits Systems Video Technology, Vol. 8, No. 5, pp. 602-615, Sep. 1998.

[30] Chang, S.-F., J.R. Smith, M. Beigi, and A.B. Benitez, "Visual Information Retrieval from Large Distributed On-line Repositories", Communications of the ACM, Vol. 40, No. 12, pp. 63-71, Dec. 1997.

[31] Chen, J., C.A. Bouman, and J.C. Dalton, "Hierarchical Browsing and Search of Large image Databases", IEEE Trans. on Image Processing, Vol. 9, No. 3, 2000.

[32] Chen, J., C.A. Bouman, and J.C. Dalton, "Similarity Pyramids for Browsing and Organization of Large Image Databases", in Proc. SPIE/IS&I Conf. Human Vision Electronic Imaging III, Vol. 3299, San Jose, CA, Jan. 29, 1998.

[33] Chen, J., C.A. Bouman, and J.C. Dalton, "Active Browsing using Similarity Pyramid", in Proc. SPIE/IS&I Conf. on Storage and Retrieval for Image and Video Databases VII, Vol. 3656, pp 144-154, San Jose, CA, Jan. 1999.

[34] Chen, C., G. Gagaudakis, and P.L. Rosin, "Similarity-Based Image Browsing", in Proc. of IFIP World Computer Congress, Int. Conf. on Intelligent Information Processing, pp. 206-213, Beijing, China, 2000.

[35] Clitherow, P., D. Riecken, and M. Muller, "VISAR: A System for Inference and Navigation in Hypertext", in Proc. of ACM Conf. on Hypertext, Pittsburgh, PA USA, Nov. 5-8, 1989.

[36] Combs, T.T.A., and B.B. Bederson, "Does Zooming Improve Image Browsing?", in Proc. of the ACM Int. Conf. on Digital Libraries, pp. 130-137, 1999.

[37] Cooke, M., P. Green, L. Josifovski, and A. Vizinho, "Robust Automatic Speech Recognition with Missing and Uncertain Acoustic Data", Speech Communication, 2001.

[38] Craver, S., B.-L. Yeo, and M. Yeung, "Multi-Linearization Data Structure for Image Browsing", in Proc. SPIE/IS&I Conf. on Storage and Retrieval for Image and Video Databases VII, Vol. 3656, pp. 155-166, Jan. 1999.

[39] CYCORP, "The Cyc Knowledge Server", retrieved on Sept. 7, 2004, from http://www.cyc.com/cyc/technology/whatiscyc.

[40] CYCORP, "CycL: the Cyc Representation Language", retrieved on Sept. 7, 2004, from http://www.cyc.com/cyc/technology/technology/whatiscyc_dir/howdoescycreason#cycl.

[41] CYCORP, "Knowledge-Enhanced Searching of Captioned Information", retrieved on Sept. 7, 2004, from http://www.cyc.com/cyc/cycrandd/areasofrandd_dir/is.

[42] Deerwester, S., S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman, "Indexing by Latent Semantic Indexing", Journal of the American Society for Information Science (JASIS), Vol. 41, No. 6, pp. 391-407, 1990.

[43] van Doorn, M.G.L.M., "Thesauri and the Mirror Retrieval Model: A Cognitive Approach to Intelligent Multimedia Information Retrieval", Master Thesis, Database Group, University of Twente, Enschede, The Nederlands, July 1999.

[44] Dowling, W.L., and Harwood, D.K., "Music Cognition", Academic Press, Orlando, FL, 1986.

[45] Duda, R.O., P.E. Hart, and D.G. Stork, "Pattern Classification", John Wiley & Sons, Second Edition, United States of America, 2001.

[46] Dumais, S.T., "Improving the Retrieval of Information from External Sources", Behavior Research Methods, Instruments and Computers, Vol. 23, No. 2, pp. 229-236, 1991.

[47] Duygulu, P., K. Barnard, N. de Freitas, and D. Forsyth "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary", in Proc. of European Conf. on Computer Vision (ECCV-2002), pp IV:97-112, Copenhagen, May 27-June 2, 2002.

[48] Flickner, M., H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by Image and Video Content: The QBIC System", Computer, Vol. 28, No. 9, pp. 23-32, Sep. 1995.

[49] Forsyth, D.A., and M. Fleck, "Body Plans", in Proc. of IEEE Computer Vision and Pattern Recognition (CVPR-1997), pp. 678-683, San Juan, Puerto Rico, 1997.

[50] Fowler, R.H., A.W. Bradley, and W.A.L. Fowler, "Information Navigator: An Information System Using Associative Networks for Display and Retrieval", University of Texas - Pan American, Department of Computer Science, Technical Report NAG9-551, #92-1.

[51] Fowler, R.H., W.A. D. Fowler, and J.L. Williams, "Document Explorer Visualization of WWW Document and Term Spaces", University of Texas - Pan American, Department of Computer Science, Technical Report, NAG9-551, #96-6, 1996.

[52] Gomez-Perez, A., "Evaluation of Taxonomic Knowledge in Ontologies and Knowledge Bases", in Proc. of Workshop on Knowledge Acquisition, Modeling and Management (KAW-1999), Alberta, Canada, Oct. 16-21, 1999.

[53] Gonzalo, J., F.V. Irina, and C.J. Cigarran, "Indexing with WordNet Synsets Can Improve Text Retrieval", in Proc. COLING-ACL Workshop on Usage of WordNet in Natural Language Processing Systems, Montreal, Canada, Aug. 16, 1998.

[54] Grosky, W.I., and R. Zhao: "Negotiating the Semantic Gap: From Feature Maps to Semantic Landscapes", in Proc. of Conf. on Current Trends in Theory and Practice of Informatics (SOFSEM-2001), Piestany, Slovak Republic, Nov. 24-De. 1, 2001.

[55] Guglielmo, E.J. and N.C. Rowe, "Natural-Language Retrieval of Images Based on Descriptive Captions',' ACM Trans. on Information Systems, Vol. 14, No. 3, July 1996, pp. 237 - 267.

[56] Hastings, W.K., "Monte Carlo Sampling Methods Using Markov Chains and their Applications", Biometrika, Vol. 57, No. 1, pp. 97-109, 1970.

[57] Hocker, B., "Bill Hocker. Photographs. Albums ", retrieved on Sep. 9, 2004, from http://www.billhocker.com/albums/.

[58] Hofmann, T., and J. Puzicha "Statistical Models for Co-Occurrence Data", A.I. Memo 1635, Massachusetts Institute of Technology, 1998.

[59] Honkela, T., S. Kaski, K. Lagus, and T. Kohonen, "WEBSOM - Self-Organizing Maps of Document Collections", in Proc. of the Workshop on Self-Organizing Maps (WSOM-1997), pp. 310-315, Espoo, Finland, June 4-6, 1997.

[60] Institution of Electrical Engineers, "INSPEC Classification and Thesaurus", 1991; also retrieved on Aug 18, 2004, from http://www.iee.org/Publish/Support/Inspec/Document/Class/index.cfm and http://www.iee.org/Publish/Support/Inspec/Document/Thes/index.cfm.

[61] ISO/IEC 15938-1..8, "Information Technology - Multimedia Content Description Interface - Part 1 to 8 (MPEG-7)", 2002, 2003, 2004.

[62] Jaimes, A., A.B. Benitez, S.-F. Chang, and A.C. Loui, "Discovering Recurrent Visual Semantics in Consumer Photographs", in Proc. of IEEE Conf. on Image Processing (ICIP-2000), Invited Chapter on Special Session on Semantic Feature Extraction in Consumer Contents, Vancouver, Canada, Sep 10-13, 2000.

[63] Jaimes, A., and S.-F. Chang, "Learning Structured Visual Detectors From User Input at Multiple Levels", International Journal of Image and Graphics (IJIG), Special Issue on Image and Video Databases, Aug. 2001.

[64] Jaimes, A., and S.-F. Chang. "A Conceptual Framework for Indexing Visual Information at Multiple Levels", in Proc. of IS&T/SPIE Conf. on Internet Imaging, San Jose, CA, Jan. 2000.

[65] Jaimes, A., and J.R. Smith, "Semi-Automatic, Data-Driven Construction of Multimedia Ontologies", in Proc. of IEEE Int. Conf. on Multimedia and Expo (ICME-2003), Baltimore, MA, USA, July 6-9, 2003.

[66] Jain, A.K., M.N. Murty, and P.J. Flynn, "Data Clustering: A Review", ACM Computing Surveys, Vol. 31, No. 3, pp.264-323, Sep. 1999.

[67] Jarvis, R.A., and E.A. Patrick, "Clustering Using a Similarity Measure Based on Shared Near Neighbors", IEEE Transaction on Computers, Vol. c-22, No. 11, Nov. 1973.

[68] Jiang, J.J., and D.W. Conrath, "Semantic Similarity based on Corpus Statistics and Lexical Taxonomy", in Proc. of Int. Conf. on Research in Computational Linguistics, Taiwan, 1997.

[69] Jing, Y., and W.B. Croft, "An Association Thesaurus for Information Retrieval", in Proc. of the RIAO Conf. on Intelligent Multimedia Information Retrieval Systems and Management (RIAO-1994), pp. 146-160, New York, NY, USA, Oct 1994.

[70] Johnson-Laird, P.N., "Mental Models", Cambridge University Press, Cambridge, MA, 1983.

[71] Jones, K.S., and P. Willett, Readings in Information Retrieval, Morgan Kaufmann Publishers, San Francisco, California, USA, 1997.

[72] Jose, J.M., J. Furner, and D.J. Harper, "Spatial Querying for Image Retrieval: a User-Oriented Evaluation", in Proc. of Int ACM SIGIR Conf on Research and Development in Information Retrieval, pp. 232-240, 1998.

[73] Jörgensen, C, "Attributes of Images in Describing Tasks", Information Processing & Management, Vol. 34, No. 2/3, 1998.

[74] Joyce, D.W., P.H. Lewis, R.H. Tansley, M.R. Dobie, and W. Hall, "Semiotics and Agents for Integrating and Navigating Through Media Representations of Concepts", in Proc. of SPIE/IS&T Conf. on Storage and Retrieval for Media Databases, Vol. 3972, pp.120-31, San Jose, CA, USA, Jan. 2000.

[75] Kamada, T., and S. Kawai, "An Algorithm for Drawing General Undirected Graphs", Information Processing Letter, Vol. 31, No. 1, pp. 7-15, April, 1989.

[76] Kamps, J., "Visualizing WordNet structure", in Proc. of Int. Conf. on Global WordNet (CIIL-2002), pp. 182-186, Mysore, India, 2002.

[77] Kohonen, T., "Self-Organizing Maps", Series in Information Sciences, Vol. 30, Springer-Verlag, Heidelberg, 1995; second ed. 1997.

[78] Kohonen, T., "Self-Organization of Very Large Document Collections: State of the Art", in Proc. of Conf. on Artificial Neural Networks (ICANN-1998), Skovde, Sweden, 1998.

[79] Kosslyn, S.M., "Image and Mind", Harvard University Press, Cambridge, MA, 1980.

[80] Kumar, R., and S.-F. Chang, "Image Retrieval with Sketches and Coherent Images", in Proc. of Int. Conf. on Multimedia and Expo (ICME-2000), New York, NY, Aug. 2000.

[81] Laaksonen, J., M. Koskela, S. Laakso, and E. Oja, "Self-Organizing Maps as a Relevance Feedback Technique in Content-Based Image Retrieval", Pattern Analysis and Applications, Vol. 2, No. 4, pp. 140-152, 2000.

[82] Lenat, D., "The Dimensions of Context Space", retrieved on Aug. 15, 2000, from http://www.cyc.com/context-space.doc; also retrieved on Sept. 7, 2004, from http://www.ai.mit.edu/people/phw/6xxx/lenat2.pdf.

[83] Lewis, P., H. Davis, M. Dobie, and W. Hall, "Towards Multimedia Thesaurus Support for Media-Based Navigation", in Proc. of Int. Workshop on Image Databases and Multimedia Search (IDB-MMS-1996), pp. 83-90, Amsterdam, The Nederlands, 1996.

[84] Lin, X., "Map Displays of Information Retrieval", Journal of the American Society for Information Science (JASIS), Vol. 40, No. 1, pp. 40-54, 1997.

[85] Ma, W.Y., and B.S. Manjunath, "A Texture Thesaurus for Browsing Large Aerial Photographs", Journal of the American Society for Information Science (JASIS), Vol. 49, No. 7, pp. 633-648, May 1998.

[86] Mandala, R., T. Takenobu, and T. Hozumi, "The Use of WordNet in Information Retrieval", in Proc. COLING-ACL Workshop on Usage of WordNet in Natural Language Processing Systems, Montreal, Canada, Aug. 16, 1998.

[87] Manjunath, B.S., P. Salembier, and T. Sikora, (eds), "Introduction to MPEG-7: Multimedia Content Description Interface", Wiley, 2002.

[88] Markkula, M., M. Tico, B. Sepponen, K. Nirkkonen, and E. Sormunen, "A Test Collection for the Evaluation of Content-Based Image Retrieval Algorithms – A User and Task-Based Approach", Information Retrieval, Vol. 4, No. 3/4, pp. 275-294, 2001.

[89] MacCuish, J., A. McPherson, J. Barros, and P. Kelly, "Interactive Layout Mechanisms for Image Database Retrieval", in Proc. of SPIE/IS&T Conf. on Visual Data Exploration and Analysis III, vol. 2656, pp. 104-115, San Jose, CA, Jan31-Feb 2, 1996.

[90] McIntyre, A., "Photo Gallery", retrieved on Aug. 18, 2004, from http://www.raingod.com/angus/Gallery/index.html.

[91] McKelvie, D., C. Brew and H. Thompson, "Using SGML as a basis for data-intensive NLP", in Proc. of Conf. on Applied Natural Language Processing, Washington, USA, April 1997.

[92] Mezaris, V., I. Kompatsiaris, and M.G. Strintzis, "An Ontology Approach to Object-Based Image Retrieval", In Proc. of Int. Conf. on Image Processing (ICIP-2003), Barcelona, Spain, Sep. 14-17, 2003.

[93] Meystel, A., "Semiotic Modeling and Situation Analysis: An Introduction", AdRem, Bala Cynwyd, PA, 1995.

[94] Mihalcea, R., and D. Moldovan, "Automatic Generation of a Coarse Grained WordNet", in Proc. of North American Association for Computational Linguistics (NAACL-2001), Pittsburgh, PA, June 2001.

[95] Miller, G.A., "WordNet: A Lexical Database for English", Communication of the ACM, Vol. 38, No. 11, pp. 39-41, Nov. 1995.

[96] Miller, G.A., "The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information", Physiological Review, Vol. 63, 1956.

[97] Minka, T.P., and R.W. Picard, "Interactive Learning Using a "Society of Models", Pattern Recognition, Special Issue on Image Database: Classification and Retrieval, Vo. 30, No. 4, 1996.

[98] Minsky, M., "A Framework from Representing Knowledge", The Psychology of Computer Vision, P. Winston (ed), pp. 211-277, McGraw-Hill, New York, NY, 1975.

[99] Mojsilovic, A., and J. Gomes, "Semantic Based Categorization, Browsing and Retrieval in Medical Image Databases", in Proc. of Int. Conf. Image Processing (ICIP-2002), Rochester, New York, Sept. 2002.

[100] Mori, Y., H. Takahashi, and R. Oka, "Image-to-Word Transformation based on Dividing and Vector Quantizing Images with Words", in Proc. of Int. Workshop on Multimedia Intelligent Storage and Retrieval Management (MISRM-1999), 1999.

[101] Morse, E., M. Lewis, and K.A. Olsen, "Evaluating Visualizations: Using a Taxonomic Guide", Int. Journal of Human Computer Studies, Vol. 53, pp. 637-662, 2000.

[102] MPEG Requirements Group, "MPEG-7: Context, Objectives and Technical Roadmap, V.12", ISO/IEC JTC1/SC29/WG11 MPEG99/N2861, Vancouver, July 1999.

[103] Murphy, K., "The Bayes Net Toolbox for Matlab", Computing Science and Statistics, Vol. 33, 2001.

[104] Naphade, R.M., and T.S. Huang, "A Probabilistic Framework for Semantic Video Indexing, Filtering, and Retrieval", IEEE Trans. on Multimedia, Vol. 3, No. 1, March 2001.

[105] Naphade, R.M., I.V. Kozintsev, and T.S. Huang, "A Factor Graph Framework for Semantic Indexing", IEEE Trans. on Circuits and Systems for Video Technology, Vol. 12, No. 1, Jan. 2002.

[106] Naphade, M.R., and J.R. Smith, "Learning Regional Semantic Concepts from Incomplete Annotation", In Proc. of Int. Conf. on Image Processing (ICIP-2003), Barcelona, Spain, Sep. 14-17, 2003.

[107] Ng, A.Y., M.I. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an Algorithm", in Proc. Neural Information Processing Systems (NIPS-2001), 2001.

[108] Oami, R., A.B. Benitez, S.-F. Chang, and N. Dimitrova, "Understanding and Modeling User Interest in Consumer Videos", In Proc. of IEEE Int. Conf. on Multimedia and Expo (ICME-2004), Taipei, Taiwan, June 2004.

[109] Paek, S., A.B. Benitez, and S.-F. Chang, "Self-Describing Schemes for Interoperable MPEG-7 Multimedia Content Descriptions", In Proc. of IS&T/SPIE Symposium on Electronic Imaging: Science and Technology (EI-

1999) - Visual Communications and Image Processing (VCIP-1999), San Jose, CA, Jan. 1999.

[110] Paek, S., and S.-F. Chang, "The Case for Image Classification Systems Based on Probabilistic Reasoning", in Proc. of IEEE Int. Conf. on Multimedia and Expo (ICME-2000), New York, NY, USA, July/Aug 30-2, 2000.

[111] Paek, S., 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", in Proc. of ACM SIGIR Workshop on Multimedia Indexing and Retrieval (ACM SIGIR-1999), Berkeley, CA, USA, Aug. 1999.

[112] Page, L., S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bring Order to the Web", Stanford Digital Library Working Chapter; available at http://dbpubs.stanford.edu:8090/pub/1999-66, Sept 2002.

[113] Pecenovic, Z., M.N. Do, M. Vetterli, and P. Pu, "Integrated Browsing and Searching of Large Image Collections", in Proc. of Conf. on Visual Information Systems, Lyon, France, 2000.

[114] Picard, R.W., "Toward a Visual Thesaurus", Springer Verlag Workshops in Computing (MIRO-1995), Invited Chapter, Glasgow, UK, Sep. 1995.

[115] Prints and Photographs Division, Library of Congress; "Thesaurus for Graphic Materials I/II", 1995, 2004; also retrieved on Aug 18, 2004, from http://www.loc.gov/rr/print/tgm1/ and http://www.loc.gov/rr/print/tgm2/.

[116] Quillian, M.R., "Semantic Memory", Semantic Information Processing, M. Minsky (ed), MIT Press, Cambridge, MA, 1968.

[117] Richardson, R., and A.F. Smeaton, "Using WordNet in a Knowledge-Based Approach to Information Retrieval", Working Paper, CA-0395, School of Computer Applications, Dublin City University, Dublin, Ireland, 1995.

[118] Rodden, K., W. Basalaj, D. Sinclair, and K. Wood, "Does Organization by Similarity Assist Image Browsing?", in Proc. of ACM Conf. on Human Factors in Computing Systems, Vol. 3, pp. 190-197, 2001.

[119] Rowe, N.C., "Precise and Efficient Retrieval of Captioned Images: The MARIE Project", Library Trends, Fall 1999.

[120] Rubner, Y., L.J. Guibas, and C. Tomasi, "The Earth Mover's Distance, Multi-Dimensional Scaling, and Color-Based Image Retrieval", in Proc. of DARPA Image Understanding Workshop, 1997.

[121] Rui, Y., T. Huang, and S.-F. Chang, "Image Retrieval: Current Techniques, Open Issues, and Promising Directions", Journal of Visual Communication and Image Representation, 1999.

[122] Rui, Y., T.S. Huang, M. Ortega, and S. Mehrotra, "Relevance Feedback: A Powerful Tool for Interactive Content-Based Image Retrieval", IEEE Trans. on Circuits and Video Technology, Vol. 8, No. 5, pp. 644-655, Sept. 1998.

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

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

[125] Santini, S., and R. Jain, "Interfaces for Emergent Semantics in Multimedia Databases", in Proc. of SPIE/IS&I Conf. on Storage and Retrieval for Image and Video Databases VII, Vol. 3656, pp 167-175, San Jose, CA, Jan. 1999.

[126] Schank, R., and R. Abelson, "Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures", Lawrence Erlbaum Associates, Hillsdale, NJ, USA, 1977.

[127] Sheikholeslami, G., W. Chang, and A. Zhang, "Semantic Clustering and Querying on Heterogeneous Features for Visual Data", in Proc. of ACM Int. Multimedia Conf. (ACMM-1996), Bristol, MA, USA, Sep. 12-16 1998.

[128] Shneiderman, B., "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations", in Proc. of IEEE Symposium on Visual Languages, Boulder, CO, USA, 1996. Sep. 3-6, 1996.

[129] Skinner, B.F., "Science and Human Behavior", Macmillan, New York, 1953.

[130] Slaney, M., "Mixtures of Probability Experts for Audio Retrieval and Indexing", in Proc. of IEEE Int. Conf. on Multimedia and Expo (ICME-2002), Lausanne, Switzerland, Aug. 26-29, 2002.

[131] Smeulders, A.W.M., M. Worring, S. Santini, A. Gupta, R. Jain, "Content-Based Image Retrieval at the End of the Early Years", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 22, No. 12, pp. 1349-1380, Dec. 2000

[132] Smith, J.R., "Quantitative Assessment of Image Retrieval Effectiveness", Journal of the American Society for Information Science and Technology (JASIST), Vol. 52, No. 11, pp 969-979, 2001.

[133] Smith, J.R., and A.B. Benitez, "Content Organization", In Introduction to MPEG 7: Multimedia Content Description Language, B. S. Manjunath, P. Salembier, T. Sikora (eds.), Chap. 10, pp. 153-162, Wiley, 2002.

[134] Smith, J.R., and A. B. Benitez, "Conceptual Modeling of Audio-Visual Data", in Proc. of IEEE Int. Conf. on Multimedia and Expo (ICME-2000), New York, NY, USA, July/Aug. 30-2, 2000.

[135] Smith, J.R., and S.-F. Chang, "VisualSEEk: A Fully Automated Content-Based Image Query System", in Proc. of ACM Int. Multimedia Conf. and Exhibition (ACM MM-1996), pp. 87-98, Boston, UK, Nov. 1996; also retrieved on Sept. 7, 2004, from http://www.ee.columbia.edu/dvmm/publications/96/smith96f.pdf.

[136] Smith, J.R., and S.-F. Chang, "Visually Searching the Web for Content", IEEE Multimedia, Vol. 4, No. 3, pp. 12-20, July 1997; also retrieved on Sept. 7, 2004, from http://www.ee.columbia.edu/dvmm/publications/97/webseek-mm-mag.pdf.

[137] Smith, J.R., A. Jaimes, C.-Y. Lin, M. Naphade, A. Natsev, and B. Tseng, "Interactive Search Fusion Methods for Video Database Retrieval", in Proc. of Int. Conf. on Image Processing (ICIP-2003), Barcelona, Spain, Sep. 14-17, 2003.

[138] Smith, J.R., C. Lin, M. Naphade, A Natsev, and B Tseng, "Multimedia Semantic Indexing using Model Vectors", in Proc. of IEEE Int. Conf. on Multimedia and Expo (ICME-2003), Baltimore, MA, USA, July 6-9, 2003.

[139] Smoliar, S.W., J.D. Baker, T. Nakayama, and L. Wilcox, "Multimedia Search: An Authoring Perspective", in Proc. of Int. Workshop on Image Databases and Multimedia Search, pp. 1-8, Amsterdam, The Netherlands, Aug. 1996.

[140] Stan, D., and I.K. Sethi, "eID: a System for Exploration of Image Databases", Information Processing and Management, Vol. 39, No. 3, 335-365, May, 2003.

[141] Stetina, J., S. Kurohashi, and M. Nagao, "General Word Sense Method Based on a Full Sentential Context", in Proc. COLING-ACL Workshop on Usage of WordNet for Natural Language Processing Systems, Montreal, Canada, July 1998.

[142] Sussna, M., "Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network", in Proc. of Conf. on Information and Knowledge Management (CIKM-1993), pp. 67-74, 1993.

[143] Szummer, M., and R. Picard, "Indoor-Outdoor Image Classification", in Proc. of IEEE Int. Workshop in Content-Based Access to Image and Video Databases (CBAIVD-1998), in conjunction with ICCV'98, Bombay, India, Jan. 1998.

[144] Tansley, R., "The Multimedia Thesaurus: Adding A Semantic Layer to Multimedia Information", Ph. D. Thesis, Computer Science, University of Southampton, Southampton, UK, Aug. 2000.

[145] Tansley, R., C. Bird, W. Hall, P. Lewis, and M. Weal, "Automatic the Linking of Content and Concept", in Proc. of ACM Int. Multimedia Conf. and Exhibition (ACM MM-2000), Los Angeles, CA, USA, Oct./Nov. 30-4, 2000.

[146] The J. Paul Getty Trust, "The Art and Architecture Thesaurus", 2000; also retrieved on Aug 18, 2004, from http://www.getty.edu/research/conducting_research/vocabularies/aat/.

[147] Tian, Q., B. Moghaddam, and T.S. Huang, "Display Optimization for Image Browsing", in Proc. of Int. Workshop on Multimedia Databases and Image Communications, Sep. 2001.

[148] Torres, R.S., C.G. Silva, C.B. Medeiros, and H.V. Rocha, "Visual Structures for Image Browsing", in Proc. of Int. Conf. on Information and Knowledge Management, pp. 49-55, New Orleans, LA, USA, Nov. 3-8, 2003.

[149] Traupman, J., and R. Wilensky, "Experiments in Improving Unsupervised Word Sense Disambiguation", CSD-03-1277, Computer Science Division, University of California, Berkeley, 2003.

[150] Tseng, B.T., C.-Y. Lin, M.R. Naphade, A. Natsev, and J.R. Smith, "Normalized Classifier Fusion for Semantic Visual Concept Detection", in Proc. of Int. Conf. on Image Processing (ICIP-2003), Barcelona, Spain, Sep. 14-17, 2003.

[151] Tsai, C.-F., K. McGarry, and J. Tait, "Image Classification Using Hybrid Neural Networks", in Proc. of ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 431-432, Toronto, Canada, July 28-Aug. 1, 2003.

[152] Turtle, H., and W. B. Croft, "Inference Networks for Document Retrieval", in Proc. of Int. Conf. on Research and Development in Information Retrieval (SIGIR'-1990), pp. 1-24, Brussels, Belgium, Sep. 5-7, 1990.

[153] Urban, J., J.M. Jose, C.J. van Rijsbergen, "An Adaptive Approach Towards Content-Based Image Retrieval", in Proc. of Int. Workshop on Content-Based Multimedia Indexing (CBMI-2003), pp. 119-126, Rennes, France, Sep. 22-24, 2003.

[154] Vailaya, A., A. Jain, and H.J. Zhang, "On Image Classification: City vs. Landscape", in Proc. of IEEE Workshop on Content-Based Access of Image and Video Libraries (CBAIVL-1998), Santa Barbara, CA, USA, June 1998.

[155] Veltkamp, R.C., and M. Tanase, "Content-Based Image Retrieval Systems: A Survey", Technical Report UU-CS-2000-34, Utrecht University, 2000; also retrieved on Sept. 7, 2004, from http://www.aa-lab.cs.uu.nl/cbirsurvey/cbir-survey/.

[156] Vendrig, J., M. Worring, and A.W.M. Smeulders, "Filtering Image Browsing: Interactive Iamge Retreival by Using Database Overviews", Multimedia Tools and Applications, Vol. 15, No. 1, pp. 83-103, Sep. 2001.

[157] W3C, "Extensible Markup Language (XML)", W3C Recommendation, retrieved on Aug 18, 2004, from http://www.w3.org/XML/.

[158] W3C, " Resource Description Framework (RDF) ", W3C Recommendation, retrieved on Aug 18, 2004, from http://www.w3.org/RDF/.

[159] W3C, "XML Schema", W3C Recommendation, retrieved on Aug 18, 2004, from http://www.w3.org/XML/Schema.

[160] W3C, "Web Ontology Language (OWL)", W3C Recommendation, retrieved on Aug 18, 2004, from http://www.w3c.org/2004/OWL/

[161] Wierzbicki, M., "Photos by Martin", retrieved on Aug. 18, 2004, from http://photosbymartin.com/.

[162] Winer, B.J., "Statistical Principles in Experimental Design", McGraw-Hill, New York, NY, 1971 (2nd ed.).

[163] Witten, I.H., and E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations", Morgan Kaufmann, Oct. 1999.

[164] Yang, M.-H., D. Kriegman, and N. Ahuja, "Detecting Faces in Images: A Survey", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 24, No. 1, pp. 34-58, 2002.

[165] Yang, J., L. Wenyin, H. Zhang, Y. Zhuang, "Thesaurus-Aided Approach for Image Browsing and Retrieval", in Proc. of IEEE Int. Conf. on Multimedia and Expo (ICME-2001), Tokyo, Aug. 2001.

[166] Yang-Pelaez, J.A., "Metrics for the Design of Visual Displays of Information", Doctor of Philosophy Thesis, Dept. of Mechanical Engineering, Massachusetts Institute of Technology, 1999.

[167] Yarowsky, D., "Unsupervised Word-sense Disambiguation Rivaling Supervised Methods", Association of Computational Linguistics, 1995.

[168] Yiu, P.R., "Image Classification Using Color Cues and Texture Orientation", Master's Thesis, Massachusetts Institute of Technology, Dept. Of EECS, 1996.

[169] Zhang, J., and J. Mostafa, "Information Retrieval by Semantic Analysis and Visualization of the Concept Space of D-Lib Magazine", D-Lib Magazine, Vol. 8, No. 10, Oct. 2002.

[170] Zhang, H.J., and D. Zhong, "A Scheme for Visual Feature Based Image Indexing", in Proc. of SPIE Conf. on Storage and Retrieval for Image and Video Databases III (IS&T/SPIE-1995), pp. 36-46, San Jose, CA, USA, Feb. 1995.

[171] Zhong, D., and S.-F. Chang, "Video Object Model and Segmentation for Content-Based Video Indexing", IEEE Int. Symposium on Circuits and Systems (ISCAS-1997), Hong Kong, 1997.

[172] Zhou, X.S., and T.S. Huang, "Relevance Feedback for Image Retrieval: a Comprehensive Review", Multimedia Systems, Vol. 8, No. 6, pp. 536-544, 2003.

[173] Zier, D., and J.-R. Ohm, "Common Datasets and Queries in MPEG-7 Color Core Experiments", ISO/IEC JTC1/SC29/WG11 MPEG99/M5060, Melbourne, Australia, Oct. 1999.

11 Appendices

11.1 MPEG-7: Multimedia Content Description Interface

The international standard MPEG-7 [61][87], also known as Multimedia Content Description Interface, has standardized tools for describing different aspects of multimedia including perceptual and semantic information. The goal of MPEG-7 is to enable advanced searching, indexing, filtering, and access of multimedia by enabling interoperability among devices and applications that deal with multimedia descriptions [102]. MPEG-7 has been an international standard since 2001. This section introduces MPEG-7 and reviews relevant MPEG-7 tools for the MediaNet framework that describe structured collections, the structure, and the semantics of multimedia. We contributed to the development of these description tools within the MPEG-7 standard, among others [17][20][109].

11.1.1 Overview of MPEG-7

MPEG-7 has standardized tools for describing multimedia. The scope of the MPEG-7 standard is to define the representation of descriptions, which is the syntax and the semantics of description tools used to create multimedia descriptions. For most

description tools, the standard does not provide normative tools for either the generation or for the consumption of the description.

Overall, the standard specifies four types of normative elements: Descriptors, Description Schemes (DSs), a Description Definition Language (DDL), and coding schemes. In MPEG-7, a Descriptor (D) defines the syntax and the semantics of an elementary feature. A descriptor can deal with low-level features, which represent the signal characteristics, such as color, shape, motion, or audio energy as well as high-level features such as the title or the author. In general, the description of a piece of multimedia involves a large number of descriptors. The descriptors are structured and related within a common framework based on Description Schemes (DSs). The DSs define a model of the description using the descriptors as building blocks. In MPEG-7, the syntax of descriptors and description schemes is defined by the Description Definition Language (DDL), which is an extension of the XML Schema language [159]. The DDL is used not only to define the syntax of MPEG-7 description tools but also to allow the declaration of new description tools that are related to specific applications. Finally, the MPEG-7 Coding Schemes are designed for compressing the MPEG-7 textual XML [157] descriptions in order to satisfy application requirements for compression efficiency, error resilience, and random access, among others.

The specification of MPEG-7 is divided into different parts. The audio and video parts define descriptors and description schemes for audio data (e.g., timbre) and video data (e.g., color layout), respectively. On the other hand, the Multimedia Description Schemes

(MDSs) specifies generic descriptors and descriptor schemes for any media. MPEG-7 includes MDS tools that describe management (e.g., creator and format), content organization (e.g., collections, structured collections, and models), summaries (e.g., hierarchies of key frames), and, even, user preferences (e.g., for searching) of multimedia. Other parts of MPEG-7 address the transmission and encoding of MPEG-7 descriptions, the description definition language (DDL), the reference implementation of the standard, and the procedures for testing conformance to MPEG-7.

In the following sections we focus on the MPEG-7 structured collection, structure, and semantic description tools, which we use to encode and represent multimedia concept networks from the MediaNet framework, as we discuss in section 3.5.3.

11.1.2 Structured Collection Description Tools

The structured collection description tools [11][24][133] can describe multimedia clusters together with their attributes and the relationships among them.

Clusters represent collections of multimedia, segments and/or descriptors, among others, for example a collection of images and image regions. Textual labels (e.g., "car"), semantics (e.g., objects and events), and descriptor statistics (e.g., the mean of color descriptors) can be described for the clusters. General relationships among clusters can be specified using various standard spatial, temporal, and semantic relationships. The structured collection tools are part of the content organization tools that describe

collections (e.g., songs on an album) and models (e.g., state transition models) of multimedia.

Figure 11.1 shows an example of how the structured collection description tools can be used to describe a collection of images. In this example, the images are grouped in two clusters corresponding to outdoor images (CM2) and indoor images (CM1) in the collection. Various attributes dealing with labels, content collections, and probability models are described for each cluster. The figure also exemplifies a relationship that describes that the two image clusters do not overlap. The XML description of this example is included in Table 11.1.

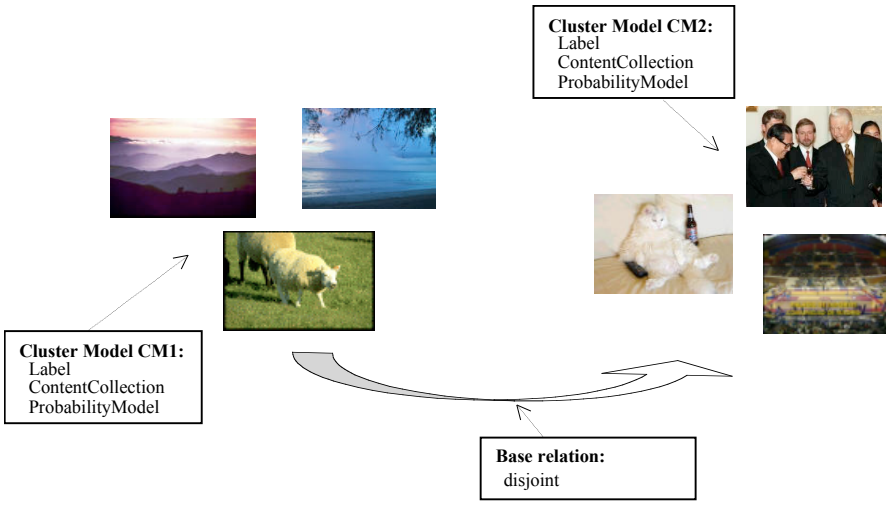


Figure 11.1: Illustration of the description of a collection of images grouped into two clusters for outdoor images (CM1) and indoor images (CM2) using MPEG-7 structured collection description tools. The XML description of this example is included in Table 11.1.

Table 11.1: The XML for the structured collection description in Figure 11.1.

```
<StructuredCollection>
 <!-- Describes a cluster with three outdoors images -->
 <ClusterModel confidence="0.9" reliability="0.9" id="CM1">
  <Label>
   <Name> Outdoors </Name>
  </Label>
  <Collection id="collection1" xsi:type="ContentCollectionType">
   <Content xsi:type="ImageType">
    <Image>
     <MediaLocator xsi:type="ImageLocatorType">
      <MediaUri>lake.jpg</MediaUri>
     </MediaLocator>
    </Image>
   </Content>
   <Content xsi:type="ImageType">
    <Image>
     <MediaLocator xsi:type="ImageLocatorType">
      <MediaUri>mountains.jpg</MediaUri>
     </MediaLocator>
    </Image>
   </Content>
   <Content xsi:type="ImageType">
    <Image>
     <MediaLocator xsi:type="ImageLocatorType">
      <MediaUri>sheeps.jpg</MediaUri>
     </MediaLocator>
    </Image>
   </Content>
```

```
  </Collection>
  <DescriptorModel>
   <Descriptor xsi:type="ScalableColorType" numOfCoeff="16"
          numOfBitplanesDiscarded="0">
    <Coeff> 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 </Coeff>
   </Descriptor>
   <Field>Coeff</Field>
  </DescriptorModel>
  <ProbabilityModel xsi:type="ProbabilityDistributionType" confidence="1.0"
          dim="16">
   <Mean mpeg7:dim="16"> 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 </Mean>
   <Variance mpeg7:dim="16"> 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 5 </Variance>
  </ProbabilityModel>
</ClusterModel>
<!-- Describes a cluster with three indoors images -->
<ClusterModel confidence="0.9" reliability="0.9" id="CM2">
  <Label>
   <Name> Indoors </Name>
  </Label>
  <Collection id="collection1" xsi:type="ContentCollectionType">
   <Content xsi:type="ImageType">
    <Image>
     <MediaLocator xsi:type="ImageLocatorType">
      <MediaUri>politics.jpg</MediaUri>
     </MediaLocator>
    </Image>
   </Content>
   <Content xsi:type="ImageType">
    <Image>
```

```
        <MediaLocator xsi:type="ImageLocatorType">
          <MediaUri>cat.jpg</MediaUri>
        </MediaLocator>
      </Image>
    </Content>
    <Content xsi:type="ImageType">
      <Image>
        <MediaLocator xsi:type="ImageLocatorType">
          <MediaUri>basketball.jpg</MediaUri>
        </MediaLocator>
      </Image>
    </Content>
  </Collection>
  <DescriptorModel>
    <Descriptor xsi:type="ScalableColorType" numOfCoeff="16"
          numOfBitplanesDiscarded="0">
      <Coeff> 6 7 8 9 0 1 2 3 4 5 6 1 2 3 4 5 </Coeff>
    </Descriptor>
    <Field>Coeff</Field>
  </DescriptorModel>
  <ProbabilityModel xsi:type="ProbabilityDistributionType" confidence="1.0"
          dim="16">
    <Mean mpeg7:dim="16"> 5 6 7 8 9 0 1 2 3 4 5 0 1 2 3 4 </Mean>
    <Variance mpeg7:dim="16"> 7 8 9 0 1 2 3 4 5 6 5 2 3 4 5 6 </Variance>
  </ProbabilityModel>
</ClusterModel>
<Relationships>
  <Node id="source" href="#CM1"/>
  <Node id="target" href="#CM2"/>
```

```
  <Relation type="urn:mpeg:mpeg7:cs:BaseRelationCS:2001:disjoint"
          source="#source" target="#target"/>
 </Relationships>
</StructuredCollection>
```

11.1.3 Structure Description Tools

The structure description tools [11][19][24] represent the structure of multimedia in space and/or time by describing general and application-specific segments of multimedia together with their attributes, hierarchical decompositions, and relationships.

Segments are whole multimedia (e.g., an image) or the result of a spatial, temporal, or spatio-temporal partitioning of multimedia (e.g., an image region). Decomposition efficiently represents segment hierarchies and can be used to create tables of contents or indexes. More general graph representations are handled by various standard spatial and temporal relationships. A segment can be described by a large number of features ranging from those targeting the life cycle of the multimedia (e.g., creation and usage) to those addressing the signal characteristics such as audio, color, shape or motion properties.

Figure 11.2 shows as example of how the structure description tools can be used to describe an image. In this example, the entire image is described as a still region (SR1), which is a group of pixels in a 2D image or a video frame. This figure also exemplifies

the spatial decomposition of entire image (SR1) into two still regions (SR2 and SR3), and the description of the spatial relationship "left" between two still regions (SR2 and SR3). Various properties dealing with creation information, textual annotation, shape features, or color features are described for each still region. The XML description of this example is included in Table 11.2.

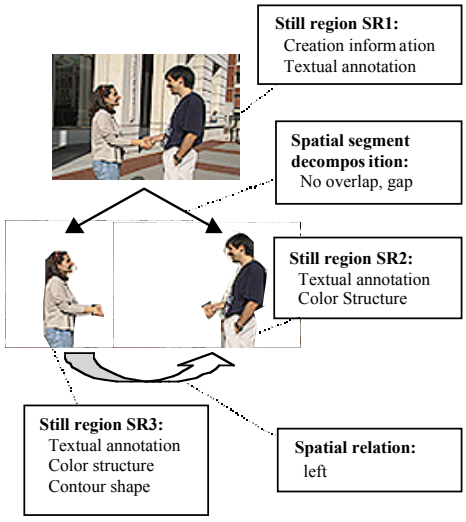


Figure 11.2: Illustration of the description of an image (SR1) and two of its regions (SR2 and SR3) using MPEG-7 structure description tools. The XML description of this example is included in Table 11.2.

Table 11.2: The XML for the structure description in Figure 11.2.

```
<StillRegion id="SR1">
 <MediaLocator>
  <MediaUri>
   http://www.ee.columbia.edu/~ana/alex&ana.jpg
  </MediaUri>
 </MediaLocator>
 <CreationInformation>
```

```
  <Creation>
   <Creator>
    <Role><Name xml:lang="en">Photographer</Name></Role>
    <Agent xsi:type="PersonType">
     <Name>
      <GivenName>Seungyup</GivenName>
     </Name>
    </Agent>
   </Creator>
   <CreationCoordinates>
    <Location>
     <Name xml:lang="en">Columbia University</Name>
     <Region>us</Region>
    </Location>
    <Date>
     <TimePoint>1998-09-19</TimePoint>
    </Date>
   </CreationCoordinates>
  </Creation>
</CreationInformation>
<TextAnnotation>
  <FreeTextAnnotation>
   Alex shakes hands with Ana
  </FreeTextAnnotation>
</TextAnnotation>
<SpatialDecomposition overlap="false" gap="true">
  <StillRegion id="SR2">
   <TextAnnotation>
    <FreeTextAnnotation> Alex </FreeTextAnnotation>
```

```
    </TextAnnotation>
    <VisualDescriptor xsi:type="ColorStructureType">
      <!-- more elements here -->
    </VisualDescriptor>
  </StillRegion>
  <StillRegion id="SR3">
    <TextAnnotation>
      <FreeTextAnnotation> Ana </FreeTextAnnotation>
    </TextAnnotation>
    <Relation type="urn:mpeg:mpeg7:cs:SpatialRelationCS:2001:left"
            target="#SR3"/>
    <VisualDescriptor xsi:type="ColorStructureType">
      <!-- more elements here -->
    </VisualDescriptor>
    <VisualDescriptor xsi:type="ContourShapeType">
      <!-- more elements here -->
    </VisualDescriptor>
  </StillRegion>
 </SpatialDecomposition>
</StillRegion>
```

11.1.4 Semantic Description Tools

The semantic description tools [11][19][21][24] represent semantic entities in narrative worlds depicted by multimedia together with their attributes and the relationships among them.

These tools can be best understood by analyzing how semantic descriptions of anything are constructed in general. One way to describe semantics is to start with events, understood as occasions when something happens. Objects, people, and places can populate such occasions and the times at which they occur. Some descriptions may concentrate on the latter aspects; with or without events, and some may even begin with a place, or an object, and describe numerous related events. Furthermore, these entities can have properties and states through which they pass as what is being described transpires. There are the interrelations among these entities. Finally, there is the world in which all of this is going on, the background, the other events and other entities, which provide context for the description and is narrative in quality. The components of the MPEG-7 semantic descriptions, therefore, fall into entities that populate narrative worlds, their attributes, and the relationships among them.

Figure 11.3 shows an example of how the semantic description tools can be used to describe the semantics of an image. This figure exemplifies the following semantic description of the image: "Alex is shaking hand with Ana, which is a symbol of comradeship". In this example, the two persons (Alex, AO1, and Ana, AO2), the event (Shake hands, EV1), and the concept (Comradeship, C1) depicted in the image are described using semantic entity description tools. The figure also exemplifies the relationships among these semantic entities, i.e., the agent, the accompanier, and the property of the event. Various attributes dealing with label, property, person, time, or

place information are described for each semantic entity. The XML description of this example is included in Table 11.3.

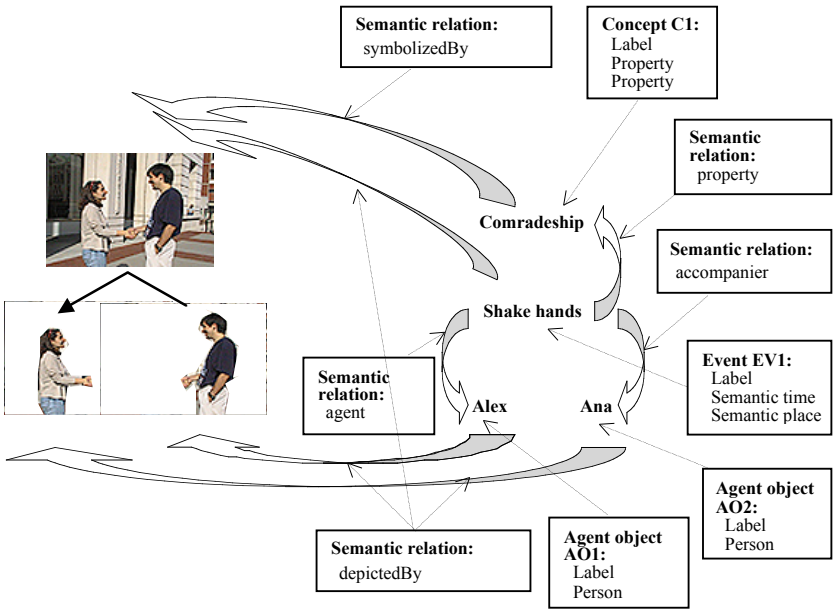


Figure 11.3: Illustration of the description of the semantics of an image using MPEG-7 semantic description tools. The example illustrates the following semantic description of the image: "Alex (AO1) is shaking hands (EV1) with Ana (AO2), which is a symbol of comradeship (C1)". The XML description of this example is included in Table 11.3.

Table 11.3: The XML for the semantic description in Figure 11.3.

```
<Semantic>
 <Label><Name>Alex shakes hands with Ana </Name></Label>
 <SemanticBase xsi:type="EventType" id="EV1">
  <Label><Name>Shake hands</Name></Label>
  <Relation type="urn:mpeg:mpeg7:cs:SemanticRelationCS:2001:agent"
        target="#AO1"/>
  <Relation type="urn:mpeg:mpeg7:cs:SemanticRelationCS:2001:accompanier"
        target="#AO2"/>
```

```
  <Relation type="urn:mpeg:mpeg7:cs:SemanticRelationCS:2001:property"
        target="#C1"/>
  <Relation type="urn:mpeg:mpeg7:cs:SemanticRelationCS:2001:depictedBy"
        target="#SR1"/>
  <SemanticPlace>
   <Label><Name>Columbia University</Name></Label>
  </SemanticPlace>
  <SemanticTime>
   <Label><Name>September 9, 1998</Name></Label>
   <Time>
    <TimePoint>1998-09-09T09</TimePoint>
   </Time>
  </SemanticTime>
</SemanticBase>
<SemanticBase xsi:type="AgentObjectType" id="AO1">
  <Label><Name>Alex</Name></Label>
  <Relation type="urn:mpeg:mpeg7:cs:SemanticRelationCS:2001:depictedBy"
        target="#SR2"/>
  <Agent xsi:type="PersonType">
   <Name><GivenName>Alex</GivenName></Name>
  </Agent>
</SemanticBase>
<SemanticBase xsi:type="AgentObjectType" id="AO2">
  <Label><Name>Ana</Name></Label>
  <Relation type="urn:mpeg:mpeg7:cs:SemanticRelationCS:2001:depictedBy"
        target="#SR3"/>
  <Agent xsi:type="PersonType">
   <Name><GivenName>Ana</GivenName></Name>
  </Agent>
```

```
  </SemanticBase>
  <SemanticBase xsi:type="ConceptType" id="C1">
    <Label><Name>Comradeship</Name></Label>
    <Property>Associate</Property>
    <Property>Friend</Property>
    <Relation type="urn:mpeg:mpeg7:cs:SemanticRelationCS:2001:symbolizedBy"
            target="#SR1"/>
  </SemanticBase>
</Semantic>
```

11.2 Evaluation Questionnaire of Knowledge-Based Image Browsing Techniques

Evaluation Questionnaire

Browsing Techniques in IMKA

2004

I) INITIAL QUESTIONNAIRE

Name:

Familiar with computers: Yes / No / Somewhat

Familiar with image retrieval: Yes / No / Somewhat

Date:

Comments:

II) TASK DESCRIPTION

Imagine that you are a travel agent. One of your responsibilities is to design pamphlets for various travel destinations. These pamphlets consist of a body of text interspersed with 5 appropriate images.

Your task is to select relevant images for two pamphlets from two image collections. To perform this task, you can make use of two computerized image-browsing systems, the operation of which will be demonstrated to you.

II.1) SYSTEM 1

II.1.1) DEMONSTRATION OF THE SYSTEM

II.1.2) TASK 1

DESCRIPTION: Familiarize yourself with the system and the image collection. Browse the image collection for a few minutes to get a sense of the content of the images in the collection.

ACTION: Write down 5 words that, you think, better describe the content of the images in the collection.

1.

2.

3.

4.

5.

ACTION: Select 5 words from the table that, you think, better describe the content of the images in the collection.

mountain peak	bus	balcony	England	Asia
temple	bridge	person	mama	home
dancer	tourist	daiquiri	road	wind
ocean	heat	event	river	country

II.2.3) TASK 2

DESCRIPTION: Find relevant images for a pamphlet titled "Buddhist Vacations". Some of the text in the pamphlet is included below.

" … Buddhism is a religion and philosophy based on the teachings of Gautama Siddhartha, who lived between approximately 563 and 483 BCE. This religion originated in India and gradually spread throughout Asia, to Tibet, Nepal, China, Vietnam, Laos, and Japan. Buddhist monks live a secluded life and meditate in temples where they seek the truth about existence. Traditionally, monks shave their heads, wear saffron robes, and have to beg for their food. … "

ACTION: List below the identifiers of the 5 most relevant images you found.

1.

2.

3.

4.

5.

ACTION: Score the following aspects in executing the task using the scale below.

1 = strongly agree	5 = mildly disagree
2 = agree	6 = disagree
3 = mildly agree	7 = strongly disagree

4 = undecided (avoid using this!)

The task was clearly defined

1 2 3 4 5 6 7

The task was simple to execute

1 2 3 4 5 6 7

I have performed similar tasks in my life/profession

1 2 3 4 5 6 7

The browsing system was easy to use (e.g., naming, organization and display of image folders, browsing options)

1 2 3 4 5 6 7

The browsing system was intuitive

1 2 3 4 5 6 7

The browsing system was useful in performing the task

1 2 3 4 5 6 7

The browsing system was stimulating for performing the task

1 2 3 4 5 6 7

I felt in control using the system to browse for images

1 2 3 4 5 6 7

I felt comfortable using the browsing system

1 2 3 4 5 6 7

I had an idea of the images that would satisfy the task before starting to browse

1 2 3 4 5 6 7

The retrieved images match my initial idea very closely

1 2 3 4 5 6 7

I frequently changed my mind on he images that I was looking for

1 2 3 4 5 6 7

Working through the image browser gave me alternate ideas

1 2 3 4 5 6 7

I am happy with the images I chose

1 2 3 4 5 6 7

I believe I have seen all the possible images in the collection that satisfy my requirements

1 2 3 4 5 6 7

I believe I have succeeded in my performance of the task

1 2 3 4 5 6 7

II.2) SYSTEM 2

II.2.1) DEMONSTRATION OF THE SYSTEM

II.2.2) TASK 1

DESCRIPTION: Familiarize yourself with the system and the image collection. Browse the image collection for a few minutes to get a sense of the content of the images in the collection.

ACTION: Write down 5 words that, you think, better describe the content of the images in the collection.

1.

2.

3.

4.

5.

ACTION: Select 5 words from the table that, you think, better describe the content of the images in the collection.

mama	bridge	home	tourist	river
heat	country	temple	road	Asia
dancer	England	daiquiri	ocean	wind
mountain peak	person	event	balcony	bus

II.2.3) TASK 2

DESCRIPTION: Find relevant images for a pamphlet titled "Romantic Getaways". Some of the text in the pamphlet is included below.

" … Everyone has dreamed of that perfect romantic vacation, honeymoon, or getaway – warm tropical breezes, beautiful sunsets, or stunning mountain views – alone with your partner. Whether your dream includes an oceanfront cottage with strolls and picnics on the beach in the moonlight, or a mountain cabin with breathtaking sunsets on snow-capped peaks, you're sure to be inspired by these ideas. … "

ACTION: List below the identifiers of the 5 most relevant images you found.

1.

2.

3.

4.

5.

ACTION: Score the following aspects in executing the task using the scale below.

1 = strongly agree	5 = mildly disagree
2 = agree	6 = disagree
3 = mildly agree	7 = strongly disagree

4 = undecided (avoid using this!)

The task was clearly defined

1 2 3 4 5 6 7

The task was simple to execute

1 2 3 4 5 6 7

I have performed similar tasks in my life/profession

1 2 3 4 5 6 7

The browsing system was easy to use (e.g., naming, organization and display of image folders, browsing options)

1 2 3 4 5 6 7

The browsing system was intuitive

1 2 3 4 5 6 7

The browsing system was useful in performing the task

1 2 3 4 5 6 7

The browsing system was stimulating for performing the task

1 2 3 4 5 6 7

I felt in control using the system to browse for images

1 2 3 4 5 6 7

I felt comfortable using the browsing system

1 2 3 4 5 6 7

I had an idea of the images that would satisfy the task before starting to browse

1 2 3 4 5 6 7

The retrieved images match my initial idea very closely

1 2 3 4 5 6 7

I frequently changed my mind on he images that I was looking for

1 2 3 4 5 6 7

Working through the image browser gave me alternate ideas

1 2 3 4 5 6 7

I am happy with the images I chose

1 2 3 4 5 6 7

I believe I have seen all the possible images in the collection that satisfy my requirements

1 2 3 4 5 6 7

I believe I have succeeded in my performance of the task

1 2 3 4 5 6 7

II.3) WITHOUT A SYSTEM

II.3.1) TASK 1

DESCRIPTION: Look at the screen. For each image on the left, pick the image on the right that will most likely appear in a pamphlet together with the image on the left. Write the identifier of the image. In the Topic column, fill out the possible topic of the pamphlet.

1. Topic:

2. Topic:

3. Topic:

4. Topic:

5. Topic:

6. Topic:

II.5) SYSTEM FEEDBACK

TASK: Please, answer the following questions and comment on the system, the tasks, and the questionnaire of this evaluation. In particular, compare both systems you have used to browse images (e.g., naming, organization and display of image folders, browsing options).

Which system was most useful in performing the tasks?

1. System 1

2. System 2

Which system did you like best?

1. System 1

2. System 2

Please, explain your selections. You are also requested to check the coherence of these scores with the scores you individually assigned to each system in previous forms.