Preface

Content-based visual information retrieval (CBVIR) is one of the most interesting research topics in the last years for the image and video community. With the progress of electronic equipment and computer techniques for visual information capturing and processing, a huge number of image and video records have been collected. Visual information becomes a well-known information format and a popular element in all aspects of our society. The large visual data make the dynamic research focused on the problem of how efficiently to capture, store, access, process, represent, describe, query, search, and retrieve their contents. In the last years, this field has experienced significant growth and progress, resulting in a virtual explosion of published information.

The research on CBVIR has already a history of more than a dozen years. It was started by using low-level features, such as color, texture, shape, structure, and space relationship, as well as (global and local) motion to represent the information content. Research on feature-based visual information retrieval has made quite a bit, but limited, success. Due to the considerable difference between the users' concerns on the semantic meaning and the appearances described by the aforementioned low-level features, the problem of semantic gap arises. One has to shift the research toward some high levels, and especially the semantic level. So, semantic-based visual information retrieval (SBVIR) began a few years ago and soon became a notable theme of CBVIR.

Research on SBVIR is conducted around various topics, such as (in an alpha-beta list) distributed indexing and retrieval, higher-level interpretation, human-computer interaction, human knowledge and behavior, indexing and retrieval for huge databases, information mining for indexing, machine-learning approaches, news and sport video summarization, object classification and annotation, object recognition and scene understanding, photo album and

storybook generation, relevance feedback and association feedback, semantic modeling for retrieval, semantic representation and description, semiotics and symbolic operation, video abstraction and structure analysis, and so forth.

How to bridge the gap between semantic meaning and the perceptual feeling, which also exists between man and computer, has attracted much attention. Many efforts have converged to SBVIR in recent years, though it is still in its commencement. As a consequence, there is a considerable requirement for books like this one, which attempts to make a summary of the past progresses and to bring together a broad selection of the latest results from researchers involved in state-of-the-art work on semantic-based visual information retrieval.

This book is intended for scientists and engineers who are engaged in research and development of visual information (especially image and video content) techniques and who wish to keep their paces with the advances of this field. The objective of this collection is to review and survey new forward-thinking research and development in intelligent, content-based retrieval technologies. A comprehensive coverage of various branches of semantic-based visual information retrieval is provided by more than 30 leading experts around the world.

The book includes 16 chapters that are organized into six sections. They cover several distinctive research fields in semantic-based visual information retrieval and form a solid background for treating the content-based process from low-level to high-level both at static and dynamic aspects. They also provide many state-of-the-art advancements and achievements in filling the semantic gap. Some detailed descriptions for each section and chapter are provided in the following.

Section I, Introduction and Background, consists of one opening and surveying chapter (Chapter I) and provides some surrounding information, various achieved results, and a brief overview of the current research foci.

Chapter I, *Toward High-Level Visual Information Retrieval*, considers content-based visual information retrieval (CBVIR) as a new generation (with new concepts, techniques, mechanisms, etc.) of visual information retrieval, provides a general picture about research and development on this subject. The research on CBVIR starts by using low-level features more than a dozen years ago. The current focus is around capturing high-level semantics (i.e., the so-called semantic-based visual information retrieval [SBVIR]). This chapter conveys the research from feature level to semantic level by treating the problem of semantic gap under the general framework of CBVIR. This chapter first shows some statistics about the research publications on semantic-based retrieval in recent years in order to give an idea about its development status. It then presents some effective approaches for multi-level image retrieval and multi-level video retrieval. This chapter also gives an overview of several current centers of attention by summarizing certain results on those subjects, such as image and video annotation, human-computer interaction, models and tools for semantic retrieval, and miscellaneous techniques in application. In addition, some future research directions such as domain knowledge and learning, relevance feedback and association feedback, as well as research at even higher levels such as cognitive level and affective level are pointed out.

Section II, From Features to Semantics, discusses some techniques on the road. It is recognized that high-level research often relies on low-level investigation, so the development of feature-based techniques would help semantic-based techniques considerably. This section consists of three chapters (Chapters II through IV).

Chapter II, *The Impact of Low-Level Features in Semantic-Based Image Retrieval*, provides a suitable starting point for going into the complex problem of content representation and

description. Considering image retrieval (IR) as a collection of techniques for retrieving images on the basis of features (in its general sense), both the low-level (content-based IR) feature and the high-level (semantic-based IR) feature (especially their relations) are discussed. Since semantic-based features rely on low-level ones, this chapter tries to make the reader initially familiarized with the most widely used low-level features. An efficient way to present these features is by means of a statistical tool capable of bearing concrete information, such as the histogram. For use in IR, histograms extracted from the low level features need to be compared by means of a metric. The most popular methods and distance metrics are, thus, opposed. Finally, several IR systems using histograms are presented in a thorough manner, and some experimental results are discussed. The steps in order to develop a custom IR system, along with modern techniques in image feature extraction, also are presented.

Chapter III is titled *Shape-Based Image Retrieval by Alignment*. Among the existing CBIR techniques for still images based on different perceptual features, shape-based methods are particularly challenging due to the intrinsic difficulties in dealing with shape localization and recognition problems. Nevertheless, there is no doubt that shape is one of the most important perceptual features, and successful shape-based techniques would significantly improve the spreading of general-purpose image retrieval systems. In this chapter, a shape-based image retrieval approach that is able to deal efficiently with domain-independent images with possible cluttered backgrounds and partially occluded objects is proposed. It is based on an alignment approach proven to be robust in rigid object recognition, which has been modified in order to deal with inexact matching between the stylized shape input by the user as query and the real shapes represented in the system's database. Results with a database composed of complex real-life images randomly taken from the Web and composed of several objects are provided.

Chapter IV is titled *Statistical Audiovisual Data Fusion for Video Scene Segmentation*. Video has a large data volume and complex structure. Automatic video segmentation into semantic units is important in effectively organizing long videos. In this chapter, the focus is on the problem of video segmentation into narrative units called scenes-aggregates of shots unified by a common dramatic event. A statistical video scene segmentation approach that detects scenes boundaries in one pass by fusing multimodal audiovisual features in a symmetrical and scalable manner is derived. The approach deals properly with the variability of real-valued features and models their conditional dependence on the context. It also integrates prior information concerning the duration of scenes. Two kinds of features, video coherence and audio dissimilarity, extracted both in visual and audio domains are used in the process of scene segmentation. The results of experimental evaluations carried out with ground truth video are reported. They show that this approach effectively fuses multiple modalities with higher performance, compared with an alternative rule-based fusion technique.

Section III focuses on Image and Video Annotation, which gets a lot of attention from the SBVIR research community. The text could be considered as a compact medium that expresses more abstract sense than do by image and video. So, by annotating image and video with characteristic textural entities, their semantic contents would be represented efficiently and would be used in retrieval. This section consists of three chapters (Chapters V through VII).

Chapter V is titled *A Novel Framework for Image Categorization and Automatic Annotation*. Image classification and automatic annotation could be treated as effective solutions to enable keyword-based semantic image retrieval. Traditionally, image classification and

automatic annotation are investigated in different models separately. In this chapter, a novel framework combining image classification and automatic annotation by learning semantic concepts of image categories is proposed. In order to choose representative features for obtaining information from image, a feature selection strategy is proposed, and visual keywords are constructed by using both discrete and continuous methods. Based on the selected features, the integrated patch (IP) model is proposed to describe the properties of different image categories. As a generative model, the IP model describes the appearance of the mixture of visual keywords in considering the diversity of the object composition. The parameters of the IP model then are estimated by EM algorithm. Some experimental results on Corel image dataset and Getty Image Archive demonstrate that the proposed feature selection and image description model are effective in image categorization and automatic image annotation, respectively.

Chapter VI is titled *Automatic and Semi-Automatic Techniques for Image Annotation*. When retrieving images, users may find that it is easier to express the desired semantic content with keywords than with visual features. Accurate keyword retrieval can occur only when images are described completely and accurately. This can be achieved either through laborious manual effort or complex automated approaches. Current methods for automatically extracting semantic information from images can be classified into two classes. One is text-based methods, which use metadata such as ontological descriptions and/or texts associated with images to assign and/or refine annotations. Although highly specialized in domain- (context-) specific image annotations, the text-based methods are usually semi-automatic. Another is image-based methods, which focus on extracting semantic information directly and automatically from image content, though they are domain-independent and could deliver arbitrarily poor annotation performance for certain applications. The focus of this chapter is to create an awareness and understanding of research and advances in this field by introducing basic concepts and theories and then by classifying, summarizing, and describing works with a variety of solutions from the published literature. By identifying some currently unsolved problems, several suggestions for future research directions are pointed out.

Chapter VII is titled *Adaptive Metadata Generation for Integration of Visual and Semantic Information*. The principal concern of this chapter is to provide those in the visual information retrieval community with a methodology that allows them to integrate the results of content analysis of visual information (i.e., the content descriptors and their text-based representation) in order to attain the semantically precise results of keyword-based image retrieval operations. The main visual objects under discussion are images that do not have any semantic representations therein. Those images demand textual annotation of precise semantics, which is to be based on the results of automatic content analysis but not on the results of time-consuming manual annotation processes. The technical background and literature review on a variety of annotation techniques for visual information retrieval is outlined first. The proposed method and its implemented system for generating metadata or textual indexes to visual objects by using content analysis techniques that can bridge the gaps between content descriptors and textual information then are described.

Section IV is devoted to the topic of Human-Computer Interaction. Human beings play an important role in semantic level procedures. By putting human into the loop of computer routine, it is quite convenient to introduce the domain knowledge into description module and to greatly improve the performance of the retrieval system. This chapter consists of three chapters (Chapters VIII through X).

Chapter VIII is titled *Interaction Models and Relevance Feedback in Image Retrieval*. Human-computer interaction increasingly is recognized as an indispensable component of image retrieval systems. A typical form of interaction is that of relevance feedback, whereby users supply relevant information on the retrieved images. This information subsequently can be used to optimize retrieval parameters and to enhance retrieval performance. The first part of the chapter provides a comprehensive review of existing relevance feedback techniques and also discusses a number of limitations that can be addressed more successfully in a browsing framework. Browsing models form the focus of the second part of this chapter, in which the merit of hierarchical structures and networks for interactive image search are evaluated. This exposition aims to provide enough detail to enable the practitioner to implement many of the techniques and to find numerous pointers to the relevant literature otherwise.

Chapter IX, *Semi-Automatic Ground Truth Annotation for Benchmarking of Face Detection in Video*, presents a method of semi-automatic ground truth annotation for benchmarking of face detection in video. It aims to illustrate the solution to the issue in which an image processing and pattern recognition expert is able to label and annotate facial patterns in video sequences at the rate of 7,500 frames per hour. These ideas are extended to the semi-automatic face annotation methodology in which all object patterns are categorized into four classes in order to increase flexibility of evaluation results analysis. A strict guide on how to speed up manual annotation process by 30 times is presented and is illustrated with sample test video sequences that consist of more than 100,000 frames, including 950 individuals and 75,000 facial images. Experimental evaluation of face detection using ground truth data that were semi-automatically labeled demonstrates the effectiveness of the current approach for both learning and testing stages.

Chapter X is titled *An Ontology-Based Framework for Semantic Image Analysis and Retrieval*. In order to overcome the limitations of keyword- and content-based visual information access, an ontology-driven framework is developed in this chapter. Under the proposed framework, an appropriately defined ontology infrastructure is used to drive the generation of manual and automatic image annotations and to enable semantic retrieval by exploiting the formal semantics of ontology. In this way, the descriptions considered in the tedious task of manual annotation are constrained to named entities (e.g., location names, person names, etc.), since the ontology-driven analysis module can automatically generate annotations concerning common domain objects of interest (e.g., sunset, trees, sea, etc.). Experiments in the domain of outdoor images show that such an ontology-based scheme realizes efficient visual information access with respect to its semantics.

Section V is intended to present some Models and Tools for Semantic Retrieval, which continuously have been incorporated in recent years. As for other disciplines or subjects, the progress of research on SBVIR also should get support from a variety of mathematic models and technique tools. Several models and tools utilized in SBVIR thus are introduced, and some pleasing results also are presented. This section consists of three chapters (Chapters XI through XIII).

Chapter XI is titled *A Machine Learning-Based Model for Content-Based Image Retrieval*. Index is an important component of a retrieval system. A suitable index makes it possible to group data according to similarity criteria. Traditional index structures frequently are based on trees and use the k-nearest neighbors (k-NN) approach to retrieve databases. Due to some disadvantages of such an approach, the use of neighborhood graphs was proposed. This approach is interesting, but it also has some disadvantages consisting mainly in its complexity. This chapter first proposes an effective method for locally updating neighborhood graphs that

constitute the required index. This structure then is exploited in order to make the retrieval process using queries in an image form more easy and effective. In addition, the indexing structure is used to annotate images in order to describe their semantics. The proposed approach is based on an intelligent manner for locating points in a multidimensional space. Promising results are obtained after experimentations on various databases.

Chapter XII, *Neural Networks for Content-Based Image Retrieval*, introduces the use of neural networks for content-based image retrieval (CBIR) systems. It presents a critical literature review of both the traditional and neural network-based techniques that are used in retrieving images based on their content. It shows how neural networks and fuzzy logic can be used in various retrieval tasks, such as interpretation of queries, feature extraction, and classification of features by describing a detailed research methodology. It investigates a neural network-based technique in conjunction with fuzzy logic in order to improve the overall performance of the CBIR systems. The results of the investigation on a benchmark database with a comparative analysis are presented in this chapter. The methodologies and results presented in this chapter will allow researchers to improve and compare their methods and also will allow system developers to understand and implement the neural network and fuzzy logic-based techniques for content-based image retrieval.

Chapter XIII is titled *Semantic-Based Video Scene Retrieval Using Evolutionary Computing*. A new emotion-based video scene retrieval method is proposed in this chapter. Five features extracted from a video are represented in a genetic chromosome, and target videos that users have in mind are retrieved by the interactive genetic algorithm through the feedback iteration. After selecting the videos that contain the corresponding emotion from the initial population of videos by the proposed algorithm, the feature vectors extracted from them are regarded as chromosomes, and a genetic crossover is applied to those feature vectors. Next, new chromosomes after crossover and feature vectors in the database videos are compared based on a similarity function in order to obtain the most similar videos as solutions of the next generation. By iterating this process, a new population of videos that users have in mind is retrieved. In order to show the validity of the proposed method, six example categories—action, excitement, suspense, quietness, relaxation, and happiness—are used as emotions for experiments. This method of retrieval shows 70% of effectiveness on the average of more than 300 commercial videos.

Section VI brings together several Miscellaneous Techniques in Applications of semantic retrieval. Research on SBVIR is still in its infancy, a large number of special ideas and exceptional techniques have been applied and also are going to apply to SBVIR. These works provide new sights and fresh views from various sides and enrich the technique pool for treating the process on semantics. This chapter consists of three chapters (Chapters XIV through XVI).

Chapter XIV, *Managing Uncertainties in Image Databases*, focuses on those functionalities of multimedia databases that are not present in traditional databases but are needed when dealing with multimedia information. Multimedia data are inherently subjective; for example, the association of a meaning and the corresponding content description to an image as well as the evaluation of the difference between two images usually depend on the user, who is involved in the evaluation process. For retrieval purposes, such subjective information needs to be combined with objective information such as image color histograms obtained through (generally imprecise) data analysis processes. Therefore, the inherently fuzzy nature of multimedia data, at both the subjective and objective sides, may lead to multiple, possibly inconsistent interpretations of data. In this chapter, a fuzzy, nonfirst, normal form

(FNF^2) data model that is an extension of the relational models is presented. It takes into account subjectivity and fuzziness. It is intuitive and enables user-friendly information access and manipulation mechanisms. A prototype system based on the FNF2 model has been implemented.

Chapter XV is titled *A Hierarchical Classification Technique for Semantics-Based Image Retrieval*. A new approach with multiple steps for improving image retrieval accuracy by integrating semantic concepts is presented in this chapter. First, images are represented according to different abstraction levels. At the lowest level, they are represented with visual features. At the upper level, they are represented with a set of very specific keywords. At the subsequent higher levels, they are represented with more general keywords. Second, visual content together with keywords are used to create a hierarchical index. A probabilistic classification approach is proposed, which allows grouping similar images into the same class. Finally, this index is exploited to define three retrieval mechanisms: text-based, content-based, and a combination of both. Experiments show that such a combination allows nicely narrowing the semantic gap encountered by most current image retrieval systems. Furthermore, it is shown that the proposed method helps to reduce retrieval time and improve retrieval accuracy.

Chapter XVI is titled *Semantic Multimedia Information Analysis for Retrieval Applications*. Most of the research works in multimedia retrieval applications have focused on retrieval by content or retrieval by example. Since the beginning of the century, a new interest has grown immensely in the multimedia information retrieval community: retrieval by semantics. This exciting new research area arises as a combination of multimedia understanding, information extraction, information retrieval, and digital libraries. This chapter presents a comprehensive review of analysis algorithms in order to extract semantic information from multimedia content. Some statistical approaches to analyze image and video contents are described and discussed.

Overall, the 16 chapters in six sections with hundreds of pictures and several dozen tables offer a comprehensive image about the current advancements of semantic-based visual information retrieval.

Yu-Jin Zhang

Editor

Tsinghua University, Beijing, China