57

Multimedia Data Mining

Zhongfei (Mark) Zhang and Ruofei Zhang

[1] SUNY at Binghamton, NY 13902-6000, zhongfei@cs.binghamton.edu
[2] Yahoo!, Inc., Sunnyvale, CA 94089 rzhang@yahoo-inc.com

Summary. *Each chapter should be preceded by an abstract (10–15 lines long) that summarizes the content. The abstract will appear *online* at www.SpringerLink.com and be available with unrestricted access. This allows unregistered users to read the abstract as a teaser for the complete chapter. As a general rule the abstracts will not appear in the printed version of your book unless it is the style of your particular book or that of the series to which your book belongs. Please use the 'starred' version of the new Springer abstract command for typesetting the text of the online abstracts (cf. source file of this chapter template abstract) and include them with the source files of your manuscript. Use the plain abstract command if the abstract is also to appear in the printed version of the book.

57.1 Introduction

Multimedia data mining, as the name suggests, presumably is a combination of the two emerging areas: *multimedia* and *data mining*. However, multimedia data mining is *not* a research area that just simply combines the research of multimedia and data mining together. Instead, the multimedia data mining research focuses on the theme of merging multimedia and data mining research together to exploit the synergy between the two areas to promote the understanding and to advance the development of the knowledge discovery in multimedia data. Consequently, multimedia data mining exhibits itself as a unique and distinct research area that synergistically relies on the state-of-the-art research in multimedia and data mining but at the same time fundamentally differs from either multimedia or data mining or a simple combination of the two areas.

Multimedia and data mining are two very interdisciplinary and multidisciplinary areas. Both areas started in early 1990s with only a very short history. Therefore, both areas are relatively young areas (in comparison, for example, with many well established areas in computer science such as operating systems, programming languages, and artificial intelligence). On the other hand, with substantial application demands, both areas have undergone independently and simultaneously rapid developments in recent years.

Multimedia is a very diverse, interdisciplinary, and multidisciplinary research area[3]. The word *multimedia* refers to a combination of multiple media types together. Due to the advanced

[3] Here we are only concerned with a research area; multimedia may also be referred to industries and even social or societal activities.

O. Maimon, L. Rokach (eds.), *Data Mining and Knowledge Discovery Handbook*, 2nd ed.,
DOI 10.1007/978-0-387-09823-4_57, © Springer Science+Business Media, LLC 2010

development of the computer and digital technologies in early 1990s, multimedia began to emerge as a research area (Furht, 1996, Steinmetz & Nahrstedt, 2002). As a research area, multimedia refers to the study and development of an effective and efficient multimedia system targeting a specific application. In this regard, the research in multimedia covers a very wide spectrum of subjects, ranging from multimedia indexing and retrieval, multimedia databases, multimedia networks, multimedia presentation, multimedia quality of services, multimedia usage and user study, to multimedia standards, just to name a few.

While the area of multimedia is so diverse with many different subjects, those that are related to multimedia data mining mainly include multimedia indexing and retrieval, multimedia databases, and multimedia presentation (Faloutsos et al., 1994, Jain, 1996, Subrahmanian, 1998). Today, it is well known that multimedia information is ubiquitous and is often required, if not necessarily essential, in many applications. This phenomenon has made multimedia repositories widespread and extremely large. There are tools for managing and searching within these collections, but the need for tools to extract hidden useful knowledge embedded within multimedia collections is becoming pressing and central for many decision-making applications. For example, it is highly desirable for developing the tools needed today for discovering relationships between objects or segments within images, classifying images based on their content, extracting patterns in sound, categorizing speech and music, and recognizing and tracking objects in video streams.

At the same time, researchers in multimedia information systems, in the search for techniques for improving the indexing and retrieval of multimedia information, are looking for new methods for discovering indexing information. A variety of techniques, from machine learning, statistics, databases, knowledge acquisition, data visualization, image analysis, high performance computing, and knowledge-based systems, have been used mainly as research handcraft activities. The development of multimedia databases and their query interfaces recalls again the idea of incorporating multimedia data mining methods for dynamic indexing.

On the other hand, data mining is also a very diverse, interdisciplinary, and multidisciplinary research area. The terminology *data mining* refers to knowledge discovery. Originally, this area began with knowledge discovery in databases. However, data mining research today has been advanced far beyond the area of databases (Faloutsos, 1996, Han & Kamber, 2006). This is due to the following two reasons. First, today's knowledge discovery research requires more than ever the advanced tools and theory beyond the traditional database area, noticeably mathematics, statistics, machine learning, and pattern recognition. Second, with the fast explosion of the data storage scale and the presence of multimedia data almost everywhere, it is not enough for today's knowledge discovery research to just focus on the structured data in the traditional databases; instead, it is common to see that the traditional databases have evolved into data warehouses, and the traditional structured data have evolved into more non-structured data such as imagery data, time-series data, spatial data, video data, audio data, and more general multimedia data. Adding into this complexity is the fact that in many applications these non-structured data do not even exist in a more traditional "database" anymore; they are just simply a collection of the data, even though many times people still call them databases (e.g., image database, video database).

Examples are the data collected in fields such as art, design, hypermedia and digital media production, case-based reasoning and computational modeling of creativity, including evolutionary computation, and medical multimedia data. These exotic fields use a variety of data sources and structures, interrelated by the nature of the phenomenon that these structures describe. As a result there is an increasing interest in new techniques and tools that can detect and discover patterns that lead to new knowledge in the problem domain where the data have been collected. There is also an increasing interest in the analysis of multimedia data generated by different distributed applications, such as collaborative virtual environments, virtual communities, and multi-agent systems. The data collected from such environments include a record of the actions in them, a variety of documents that are part of the business process, asynchronous threaded discussions, transcripts from synchronous communications, and other data records. These heterogeneous multimedia data records require sophisticated preprocessing, synchronization, and other transformation procedures before even moving to the analysis stage.

Consequently, with the independent and advanced developments of the two areas of multimedia and data mining, with today's explosion of the data scale and the existence of the pluralism of the data media types, it is natural to evolve into this new area called *multimedia data mining*. While it is presumably true that multimedia data mining is a combination of the research between multimedia and data mining, the research in multimedia data mining refers to the synergistic application of knowledge discovery theory and techniques in a multimedia database or collection. As a result, "inherited" from its two parent areas of multimedia and data mining, multimedia data mining by nature is also an interdisciplinary and multidisciplinary area; in addition to the two parent areas, multimedia data mining also relies on the research from many other areas, noticeably from mathematics, statistics, machine learning, computer vision, and pattern recognition. Figure 57.1 illustrates the relationships among these interconnected areas.

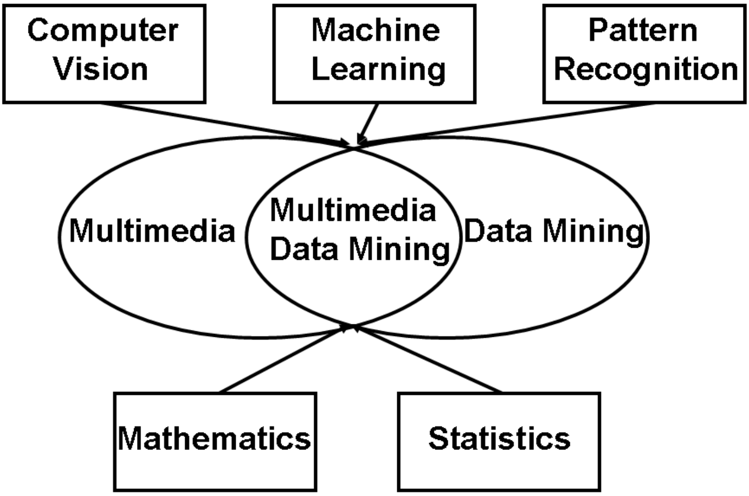


Fig. 57.1. Relationships among the interconnected areas to multimedia data mining.

While we have clearly given the working definition of multimedia data mining as an emerging, active research area, due to historic reasons, it is helpful to clarify several misconceptions and to point out several pitfalls at the beginning.

- *Multimedia Indexing and Retrieval* vs. *Multimedia Data Mining*: It is well-known that in the classic data mining research, the pure text retrieval or the classic information retrieval is *not* considered as part of data mining, as there is no knowledge discovery involved.

However, in multimedia data mining, when it comes to the scenarios of multimedia indexing and retrieval, this boundary becomes vague. The reason is that a typical multimedia indexing and/or retrieval system reported in the recent literature often contains a certain level of knowledge discovery such as feature selection, dimensionality reduction, concept discovery, as well as mapping discovery between different modalities (e.g., imagery annotation where a mapping from an image to textual words is discovered and word-to-image retrieval where a mapping from a textual word to images is discovered). In this case, multimedia information indexing and/or retrieval is considered as part of multimedia data mining. On the other hand, if a multimedia indexing or retrieval system uses a "pure" indexing system such as the text-based indexing technology employed in many commercial imagery/video/audio retrieval systems on the Web, this system is not considered as a multimedia data mining system.

- *Database* vs. *Data Collection*: In a classic database system, there is always a database management system to govern all the data in the database. This is true for the classic, structured data in the traditional databases. However, when the data become non-structured data, in particular, multimedia data, often we do not have such a management system to "govern" all the data in the collection. Typically, we simply just have a whole collection of multimedia data, and we expect to develop an indexing/retrieval system or other data mining system on top of this data collection. For historic reasons, in many literature references, we still use the terminology of "database" to refer to such a multimedia data collection, even though this is different from the traditional, structured database in concept.
- *Multimedia Data* vs. *Single Modality Data*: Although "multimedia" refers to the multiple modalities and/or multiple media types of data, conventionally in the area of multimedia, multimedia indexing and retrieval also includes the indexing and retrieval of a single, non-text modality of data, such as image indexing and retrieval, video indexing and retrieval, and audio indexing and retrieval. Consequently, in multimedia data mining, we follow this convention to include the study of any knowledge discovery dedicated to any single modality of data as part of the multimedia data mining research. Therefore, studies in image data mining, video data mining, and audio data mining alone are considered as part of the multimedia data mining area.

Multimedia data mining, although still in its early booming stage as an area that is expected to have further development, has already found enormous application potential in a wide spectrum covering almost all the sectors of society, ranging from people's daily lives to economic development to government services. This is due to the fact that in today's society almost all the real-world applications often have data with multiple modalities, from multiple sources, and in multiple formats. For example, in homeland security applications, we may need to mine data from an air traveler's credit history, traveling patterns, photo pictures, and video data from surveillance cameras in the airport. In the manufacturing domains, business processes can be improved if, for example, part drawings, part descriptions, and part flow can be mined in an integrated way instead of separately. In medicine, a disease might be predicted more accurately if the MRI (magnetic resonance imaging) imagery is mined together with other information about the patient's condition. Similarly, in bioinformatics, data are available in multiple formats.

The rest of the chapter is organized as follows. In the next section, we give the architecture for a typical multimedia data mining system or methodology in the literature. Then in order to showcase a specific multimedia data mining system and how it works, we present an example of a specific method on concept discovery in an imagery database in the following section. Finally, the chapter is concluded in Sec. 57.4.

57.2 A Typical Architecture of a Multimedia Data Mining System

A typical multimedia data mining system, or framework, or method always consists of the following three key components. Given the raw multimedia data, the very first step for mining the multimedia data is to convert a specific raw data collection (or a database) into a representation in an abstract space which is called the feature space. This process is called feature extraction. Consequently, we need a feature representation method to convert the raw multimedia data to the features in the feature space, before any mining activities are able to be conducted. This component is very important as the success of a multimedia data mining system to a large degree depends upon how good the feature representation method is. The typical feature representation methods or techniques are taken from the classic computer vision research, pattern recognition research, as well as multimedia information indexing and retrieval research in multimedia area.

Since knowledge discovery is an intelligent activity, like other types of intelligent activities, multimedia data mining requires the support of a certain level of knowledge. Therefore, the second key component is the knowledge representation, i.e., how to effectively represent the required knowledge to support the expected knowledge discovery activities in a multimedia database. The typical knowledge representation methods used in the multimedia data mining literature are directly taken from the general knowledge representation research in artificial intelligence area with the possible special consideration in the multimedia data mining problems such as spatial constraints based reasoning.

Finally, we come to the last key component — the actual mining or learning theory and/or technique to be used for the knowledge discovery in a multimedia database. In the current literature of multimedia data mining, there are mainly two paradigms of the learning or mining theory/techniques that can be used separately or jointly in a specific multimedia data mining application. They are *statistical learning theory* and *soft computing theory*, respectively. The former is based on the recent literature on machine learning and in particular statistical machine learning, whereas the latter is based on the recent literature on soft computing such as fuzzy logic theory. This component typically is the core of the multimedia data mining system.

In addition to the three key components, in many multimedia data mining systems, there are user interfaces to facilitate the communications between the users and the mining systems. Like the general data mining systems, for a typical multimedia data mining system, the quality of the final mining results can only be judged by the users. Hence, it is necessary in many cases to have a user interface to allow the communications between the users and the mining systems and the evaluations of the final mining quality; if the quality is not acceptable, the users may need to use the interface to tune different parameter values of a specific component used in the system, or even to change different components, in order to achieve better mining results, which may go into an iterative process until the users are happy with the mining results.

Figure 57.2 illustrates this typical architecture of a multimedia data mining system.

57.3 An Example — Concept Discovery in Imagery Data

In this section, as an example to showcase the research as well as the technologies developed in multimedia data mining, we address the image database modeling problem in general and, in particular, focuses on developing a hidden semantic concept discovery methodology to address effective semantics-intensive image data mining and retrieval. In the approach proposed in this section, each image in the database is segmented into regions associated with homogenous

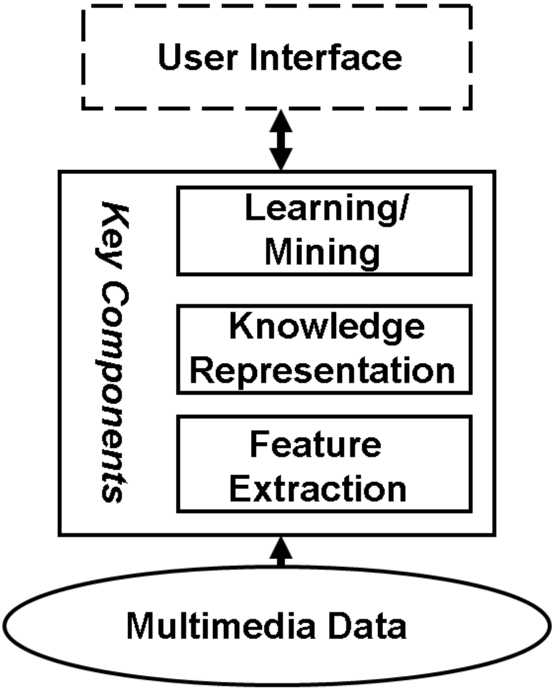


Fig. 57.2. The typical architecture of a multimedia data mining system.

color, texture, and shape features. By exploiting regional statistical information in each image and employing a vector quantization method, a uniform and sparse region-based representation is achieved. With this representation a probabilistic model based on the statistical-hidden-class assumptions of the image database is obtained, to which the Expectation-Maximization (EM) technique is applied to discover and analyze semantic concepts hidden in the database. An elaborated mining and retrieval algorithm is designed to support the probabilistic model. The semantic similarity is measured through integrating the posterior probabilities of the transformed query image, as well as a constructed negative example, to the discovered semantic concepts. The proposed approach has a solid statistical foundation; the experimental evaluations on a database of 10,000 general-purpose images demonstrate the promise and the effectiveness of the proposed approach.

57.3.1 Background and Related Work

As is obvious, large collections of images have become available to the public, from photo collections to Web pages or even video databases. To effectively mine or retrieve such a large collection of imagery data is a huge challenge. After more than a decade of research, it has been found that content based image data mining and retrieval are a practical and satisfactory solution to this challenge. At the same time, it is also well known that the performance of the existing approaches in the literature is mainly limited by the *semantic gap* between low-level features and high-level semantic concepts (Smeulders et al., 2000). In order to reduce this gap, region based features (describing object level features), instead of raw features of the whole image, to represent the visual content of an image are widely used (Carson et al., 2002, Wang et al., 2001, Jing et al., 2004, Chen & Wang, 2002).

In contrast to traditional approaches (Huang & et al., 1997, Flickner et al., 1995, Pentland et al., 1994), which compute global features of images, the region based methods extract features of the segmented regions and perform similarity comparisons at the granularity of regions. The main objective of using region features is to enhance the ability to capture and represent the focus of users' perception of the image content.

One important issue significantly affecting the success of an image data mining methodology is how to compare two images, i.e., the definition of the image similarity measurement. A straightforward solution adopted by most early systems (Carson et al., 2002, Ma & Manjunath, 1997, Wood et al., 1998) is to use individual region-to-region similarity as the basis of the comparisons. When using such schemes, the users are forced to select a limited number of regions from a query image in order to start a query session. As discussed in (Wang et al., 2001), due to the uncontrolled nature of the visual content in an image, automatically and precisely extracting image objects is still beyond the reach of the state-of-the-art in computer vision. Therefore, these systems tend to partition one object into several regions, with none of them being representative for the object. Consequently, it is often difficult for users to determine which regions should be used for their interest.

To provide users a simpler querying interface and to reduce the influence of inaccurate segmentation, several image-to-image similarity measurements that combine information from all of the regions have been proposed (Greenspan et al., 2004, Wang et al., 2001, Chen & Wang, 2002). Such systems only require users to impose a query image and therefore relieve the users from making the puzzling decisions. For example, the SIMPLIcity system (Wang et al., 2001) uses integrated region matching as its image similarity measure. By allowing a many-to-many relationship of the regions, the approach is robust to inaccurate segmentation. Greenspan et al (Greenspan et al., 2001) propose a continuous probabilistic framework for image matching. In this framework, each image is represented as a Gaussian mixture distribution, and images are compared and matched via a probabilistic measure of similarity between distributions. Improved image matching results are reported.

Ideally, what we strive to measure is the *semantic similarity*, which physically is very difficult to define, or even to describe. The majority of the existing methodologies do not explicitly connect the extracted features with the pursued semantics reflected in the visual content. They define region-to-region and/or image-to-image similarities to attempt to approximate the semantic similarity. However, the approximation is typically heuristic and consequently not reliable and effective. Thus, the retrieval and mining accuracies are rather limited.

To deal with the inaccurate approximation problem, several research efforts have been attempted to link regions to semantic concepts by supervised learning. Barnard et al proposed several statistical models (Barnard et al., 2003, Duygulu et al., 2002, Barnard & Forsyth, 2001) which connect *image blobs* and linguistic words. The objective is to predict words associated with whole images (auto-annotation) and corresponding to particular image regions (region naming). In their approaches, a number of models are developed for the joint distribution of image regions and words. The models are multi-modal and correspondence extensions to Hofmann's hierarchical clustering aspect model (Hofmann & Puzicha, 1998, Hofmann et al., 1996, Hofmann, 2001), a translation model adapted from statistical machine translation, and a multi-modal extension to the mixture of latent Dirichlet allocation models (Blei et al., 2001). The models are used to automatically annotate testing images, and the reported performance is promising. Recognizing that these models fail to exploit spatial context in the images and words, Carbonetto et al augmented the models such that spatial relationships between regions are learned. The model proposed is more expressive in the sense that the spatial correspondences are incorporated into the joint probability learning (Carbonetto et al., 2004, Carbonetto et al., 2003), which improves the accuracy of object recognition in image annotation. Recently, Feng et al proposed a Multiple Bernoulli Relevance Model (MBRM) (Feng et al., 2004) for image-word association, which is based on the Continuous-space Relevance Model (CRM) proposed by (Jeon et al., 2003). In the MBRM model, the word probabilities are estimated using a multiple Bernoulli model and the image feature probabilities using a non-parametric kernel density estimate.

We argue that for all the feature based image mining and retrieval methods, the semantic concepts related to the content of the images are always hidden. By hidden, we mean (1) objectively, there is no direct mapping from the numerical image features to the semantic meanings in the images, and (2) subjectively, given the same region, there are different corresponding semantic concepts, depending on different context and/or different user interpretations. This observation justifies the need to discover the hidden semantic concepts that is a key step toward effective image retrieval.

In this chapter, we propose a probabilistic approach to addressing the hidden semantic concept discovery. A region-based sparse but uniform image representation scheme is developed (unlike the block-based uniform representation in (Zhu et al., 2002), region-based representation is more effective for image mining and retrieval due to the fact that humans pay more attention to objects than blocks in an image), which facilitates the indexing scheme based on a region-image-concept probabilistic model with validated assumptions. This model has a solid statistical foundation and is intended for the objective of semantics-intensive image retrieval. To describe the semantic concepts hidden in the region and image distributions of a database, the Expectation-Maximization (EM) technique is used. With a derived iterative procedure, the posterior probabilities of each region in an image for the hidden semantic concepts are quantitatively obtained, which act as the basis for the *semantic similarity* measure for image mining and retrieval. Therefore, the effectiveness is improved as the similarity measure is based on the discovered semantic concepts, which are more reliable than the region features used in most of the existing systems in the literature. Figure 57.3 shows the architecture of the proposed approach.

Different from the models reviewed above, the model and the approach we propose and present here do not require training data; we formulate a generative model to discover the clusterings in a probabilistic scheme by unsupervised learning. In this model, the regions and images are connected through a hidden layer — the concept layer, which constitutes the basis of the image similarity measures. In addition, users' relevance feedback is incorporated into the model fitting procedure such that the subjectivity in image mining and retrieval is addressed explicitly and the model fitting is customized toward users' querying needs.

57.3.2 Region Based Image Representation

In the proposed approach, the query image and images in a database are first segmented into homogeneous color-texture regions. Then representative properties are extracted for every region by incorporating multiple features, specifically, color, texture, and shape properties. Based on the extracted regions, a visual token catalog is generated to explore and exploit the content similarities of the regions, which facilitates the indexing and mining scheme based on the region-image-concept probabilistic model elaborated in Section 57.3.3.

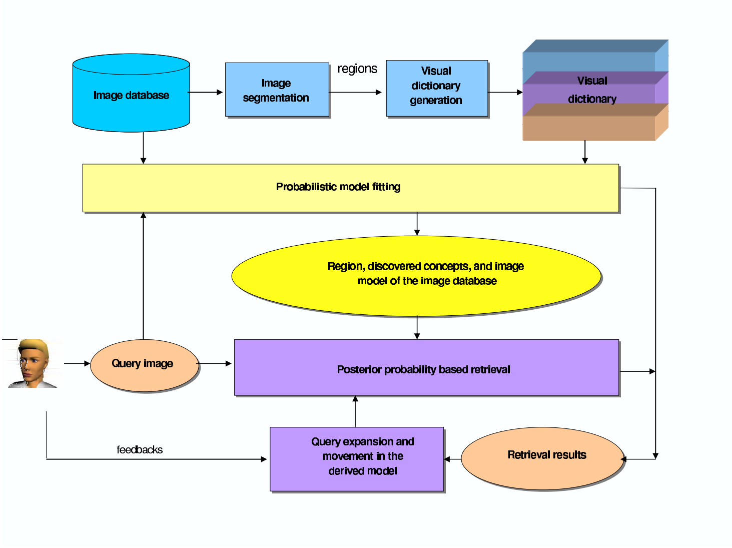


Image database
Image segmentation
regions
Visual dictionary generation
Visual dictionary
Probabilistic model fitting
Region, discovered concepts, and image model of the image database
Query image
Posterior probability based retrieval
feedbacks
Query expansion and movement in the derived model
Retrieval results

Fig. 57.3. The architecture of the latent semantic concept discovery based image data mining and retrieval approach. Reprint from (Zhang & Zhang, 2007) ©2007 IEEE Signal Processing Society Press.

Image Segmentation

To segment an image, the system first partitions the image into blocks of 4 by 4 pixels to compromise between the texture effectiveness and the computation time. Then a feature vector consisting of nine features from each block is extracted. Three of the features are average color components in the 4 by 4 pixel size block; we use the *LAB* color space due to its desired property that the perceptual color difference is proportional to the numerical difference. The other six features are the texture features extracted using wavelet analysis.

To extract texture information of each block, we apply a set of Gabor filters (Manjunath & Ma, 1996), which are shown to be effective for image indexing and retrieval (Ma & Manjunath, 1995), to the block to measure the response. The Gabor filters measure the two-dimensional wavelets. The discretization of a two-dimensional wavelet applied to the blocks is given by

$$W_{mlpq} = \int\int I(x,y)\psi_{ml}(x-p\triangle x, y-q\triangle y)dxdy \qquad (57.1)$$

where I denotes the processed block; $\triangle x$ and $\triangle y$ denote the spatial sampling rectangle; p, q are image positions; and m, l specify the scale and orientation of the wavelets. The base function $\psi_{ml}(x,y)$ is given by

$$\psi_{ml}(x,y) = a^{-m}\psi(\widetilde{x},\widetilde{y}) \qquad (57.2)$$

where

$$\widetilde{x} = a^{-m}(x\cos\theta + y\sin\theta)$$

$$\widetilde{y} = a^{-m}(-x\sin\theta + y\cos\theta)$$

denote a dilation of the mother wavelet (x,y) by a^{-m}, where a is the scale parameter, and a rotation by $\theta = l \times \triangle\theta$, where $\triangle\theta = 2\pi/V$ is the orientation sampling period; V is the number of orientation sampling intervals.

In the frequency domain, with the following Gabor function as the mother wavelet, we use this family of wavelets as our filter bank:

$$\begin{aligned}\Psi(u,v) &= \exp\{-2\pi^2(\sigma_x^2 u^2 + \sigma_y^2 v^2)\} \otimes \delta(u-W)\\ &= \exp\{-2\pi^2(\sigma_x^2 (u-W)^2 + \sigma_y^2 v^2)\}\\ &= \exp\{-\frac{1}{2}(\frac{(u-W)^2}{\sigma_u^2} + \frac{v^2}{\sigma_v^2})\}\end{aligned} \quad (57.3)$$

where $\otimes$ is a convolution symbol, $\delta(\cdot)$ is the impulse function, $\sigma_u = (2\pi\sigma_x)^{-1}$, and $\sigma_v = (2\pi\sigma_y)^{-1}$; σ_x and σ_y are the standard deviations of the filter along the x and y directions, respectively. The constant W determines the frequency bandwidth of the filters.

Applying the Gabor filter bank to the blocks, for every image pixel (p,q), in U (the number of scales in the filter bank) by V array of responses to the filter bank, we only need to retain the magnitudes of the responses:

$$F_{mlpq} = |W_{mlpq}| \quad m = 0,\ldots,U-1,\ l = 0,\ldots V-1 \quad (57.4)$$

Hence, a texture feature is represented by a vector, with each element of the vector corresponding to the energy in a specified scale and orientation sub-band w.r.t. a Gabor filter. In the implementation, a Gabor filter bank of 3 orientations and 2 scales is used for each image in the database, resulting in a 6-dimensional feature vector (i.e., 6 means for $|W_{ml}|$) for the texture representation.

After we obtain feature vectors for all blocks, we perform normalization on both color and texture features such that the effects of different feature ranges are eliminated. Then a k-means based segmentation algorithm, similar to that used in (Chen & Wang, 2002), is applied to clustering the feature vectors into several classes, with each class corresponding to one region in the segmented image.

Figure 57.4 gives four examples of the segmentation results of images in the database, which show the effectiveness of the segmentation algorithm employed.

After the segmentation, the edge map is used with the water-filling algorithm (Zhou et al., 1999) to describe the shape feature for each region due to its reported effectiveness and efficiency for image mining and retrieval (Moghaddam et al., 2001). A 6-dimensional shape feature vector is obtained for each region by incorporating the statistics defined in (Zhou et al., 1999), such as the filling time histogram and the fork count histogram. The mean of the color-texture features of all the blocks in each region is determined to combine with the corresponding shape feature as the extracted feature vector of the region.

Visual Token Catalog

Since the region features $f \in \mathbb{R}^n$, it is necessary to perform regularization on the region property set such that they can be indexed and mined efficiently. Considering that many regions from different images are very similar in terms of the features, vector quantization (VQ) techniques are required to group similar regions together. In the proposed approach, we create a visual token catalog for region properties to represent the visual content of the regions. There are three advantages to creating such a visual token catalog. First, it improves mining and retrieval robustness by tolerating minor variations among visual properties. Without the visual

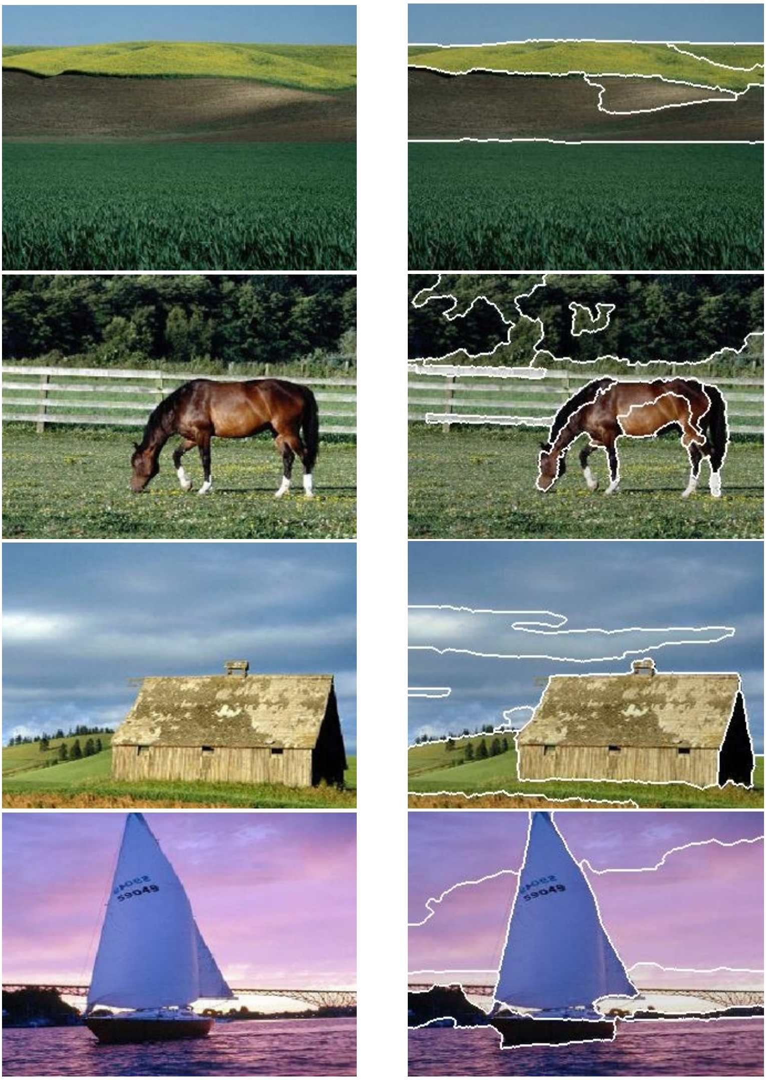


Fig. 57.4. The segmentation results. Left column shows the original images; right column shows the corresponding segmented images with the region boundary highlighted.

token catalog, since very few feature values are exactly shared by different regions, we would have to consider feature vectors of all the regions in the database. This makes it not effective to compare the similarity among regions. However, based on the visual token catalog created, low-level features of regions are quantized such that images can be represented in a way resistant to perception uncertainties (Chen & Wang, 2002). Second, the region-comparison efficiency is significantly improved by mapping the expensive numerical computation of the distances between region features to the inexpensive symbolic computation of the differences between "code words" in the visual token catalog. Third, the utilization of the visual token catalog reduces the storage space without sacrificing the accuracy.

We create the visual token catalog for region properties by applying the Self-Organization Map (SOM) (Kohonen et al., 2000) learning strategy. SOM is ideal for this problem, as it projects the high-dimensional feature vectors to a 2-dimensional plane through mapping similar features together while separating different features at the same time. The SOM learning algorithm we have used is competitive and unsupervised. The nodes in a 2-dimensional array become specifically tuned to various classes of input feature patterns in an orderly fashion.

A procedure is designed to create "code words" in the dictionary. Each "code word" represents a set of visually similar regions. The procedure follows 4 steps:

1. Performing the Batch SOM learning (Kohonen et al., 2000) algorithm on the region feature set to obtain the visualized model (node status) displayed on a 2-dimensional plane map. The distance metric used is Euclidean for its simplicity.
2. Regarding each node as a "pixel" in the 2-dimensional plane map such that the map becomes a binary lattice with the value of each pixel i defined as follows:

$$p(i) = \begin{cases} 0 \text{ if } count(i) \geq t \\ 1 \text{ else} \end{cases}$$

   where $count(i)$ is the number of features mapped to node i and the constant t is a preset threshold. Pixel value 0 denotes the objects, while pixel value 1 denotes the background.
3. Performing the morphological erosion operation (Castleman, 1996) on the resulting lattice to make sparse connected objects in the image disjointed. The size of the erosion mask is determined to be the minimum to make two sparsely connected objects separated.
4. With connected component labeling (Castleman, 1996), we assign each separated object a unique ID, a "code word". For each "code word", the mean of all the features associated with it is determined and stored. All "code words" constitute the visual token catalog to be used to represent the visual properties of the regions.

Figure 57.5 illustrates this procedure on a portion of the map we have obtained.

Simple yet effective Euclidean distance is used in the SOM learning to determine the "code word" to which each region belongs. The proof of the convergence of the SOM learning process in the 2-dimensional plane map is given in (Kohonen, 2001). The details about the selection of the parameters are also covered in (Kohonen, 2001). Each labeled component represents a region feature set among which the intra-distance is low. The extent of similarity in each "code word" is controlled by the parameters in the SOM algorithm and the threshold t. With this procedure, the number of the "code words" is adaptively determined and the similarity-based feature grouping is achieved. The experiments reported in Section 57.3.6 show that the visual token catalog created captures the clustering characteristics existing in the feature set well. We note that the threshold t is highly correlated to the number of the "code words" generated; it is determined empirically by balancing the efficiency and the accuracy.

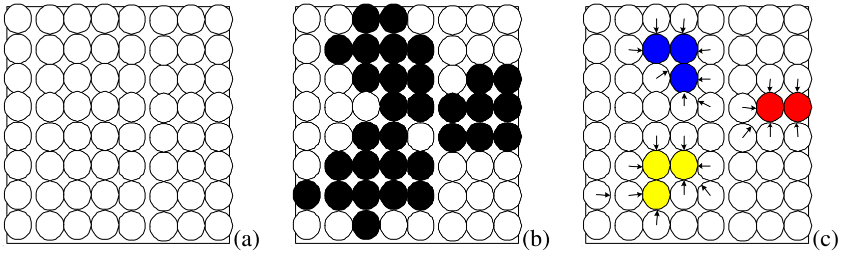


Fig. 57.5. Illustration of the procedure: (a) the initial map; (b) the binary lattice obtained after the SOM learning is converged; (c) the labeled object on the final lattice. The arrows indicate the objects that the corresponding nodes belong to. Reprint from (Zhang & Zhang, 2007) ©2007 IEEE Signal Processing Society Press.

We discuss the issue of choosing the appropriate number of the “code words” in the visual token catalog in Section 57.3.6. Figure 57.6 shows the process of the generation of the visual token catalog. Each rounded rectangle in the third column of the figure is one “code word” in the dictionary.

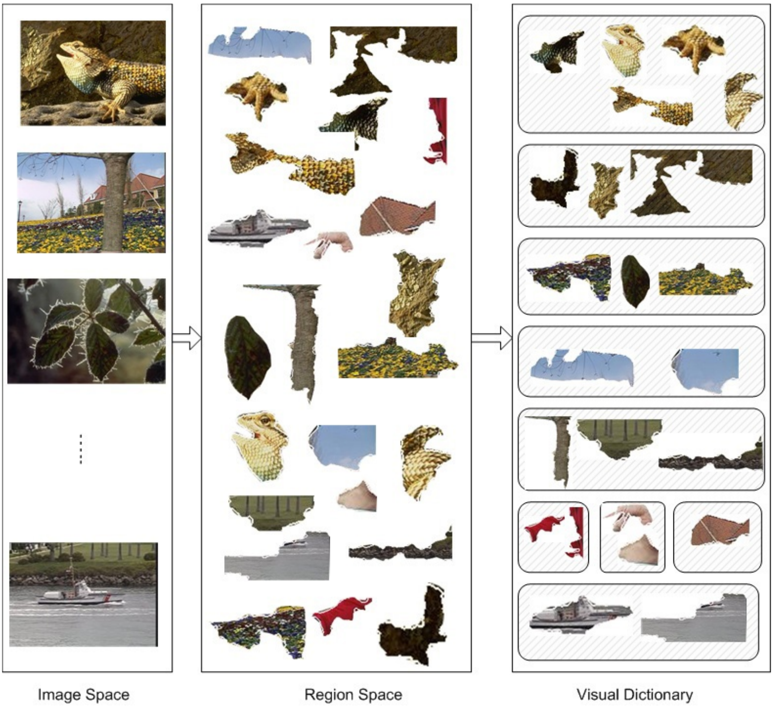


Fig. 57.6. The process of the generation of the visual token catalog. Reprint from (Zhang & Zhang, 2007) ©2007 IEEE Signal Processing Society Press and from (Zhang & Zhang, 2004a) ©2004 IEEE Computer Society Press.

For each region of an image in the database, the "code word" that it is associated with is identified and the corresponding index in the visual token catalog is stored, while the original feature of this region is discarded. For the region of a new image, the closest entry in the dictionary is found and the corresponding index is used to replace its feature. In the rest of this chapter, we use the terminologies *region* and "*code word*" interchangeably; they both denote an entry in the visual token catalog equivalently.

Based on the visual token catalog, each image is represented in a uniform vector model. In this representation, an image is a vector with each dimension corresponding to a "code word". More formally, the uniform representation $\mathbf{I}_u$ of an image I is a vector $\mathbf{I}_u = \{w_1, w_2, \dots, w_M\}$, where M is the number of the "code words" in the visual token catalog. For a "code word" $C_i, 1 \leq i \leq M$, if there exists a region R_j of I that corresponds to it, then $w_i = W_{Rj}$ for $\mathbf{I}_u$, where W_{Rj} is the number of the occurrences of R_j in the image I; otherwise, $w_i = 0$. This uniform representation is sparse, for an image usually contains a few regions compared with the number of the "code words" in the visual token catalog. Based on this representation of all the images, the database is modeled as a $M \times N$ "code word"-image matrix which records the occurrences of every "code word" in each image, where N is the number of the images in the database.

57.3.3 Probabilistic Hidden Semantic Model

To achieve the automatic semantic concept discovery, a region-based probabilistic model is constructed for the image database with the representation of the "code word"-image matrix. The probabilistic model is analyzed by the Expectation-Maximization (EM) technique (Dempster et al., 1977) to discover the latent semantic concepts, which act as a basis for effective image mining and retrieval via the concept similarities among images.

Probabilistic Database Model

With a uniform "code word" vector representation for each image in the database, we propose a probabilistic model. In this model, we assume that the specific (region, image) pairs are known i.i.d. samples from an unknown distribution. We also assume that these samples are associated with an unobserved *semantic concept* variable $z \in Z = \{z_1, \dots, z_K\}$, where K is the number of concepts to be discovered. Each observation of one region ("code word") $r \in R = \{r_1, \dots, r_M\}$ in an image $g \in G = \{g_1, \dots, g_N\}$ belongs to one concept class z_k. To simplify the model, we have two further assumptions. First, the observation pairs (r_i, g_j) are generated independently. Second, the pairs of random variables (r_i, g_j) are conditionally independent given the respective hidden concept z_k, i.e., $P(r_i, g_j|z_k) = P(r_i|z_k)P(g_j|z_k)$. Intuitively, these two assumptions are reasonable, which are further validated by the experimental evaluations. The region and image distribution may be treated as a randomized data generation process, described as follows:

- Choose a concept with probability $P(z_k)$;
- Select a region $r_i \in R$ with probability $P(r_i|z_k)$; and
- Select an image $g_j \in G$ with probability $P(g_j|z_k)$.

As a result, one obtains an observed pair (r_i, g_j), while the concept variable z_k is discarded.

Based on the theory of the generative model (Mclachlan & Basford, 1988), the above process is equivalent to the following:

- Select an image g_j with probability $P(g_j)$;

- Select a concept z_k with probability $P(z_k|g_j)$;
- Generate a region r_i with probability $P(r_i|z_k)$.

Translating this process into a joint probability model results in the expression

$$\begin{aligned} P(r_i,g_j) &= P(g_j)P(r_i|g_j) \\ &= P(g_j)\sum_{k=1}^{K} P(r_i|z_k)P(z_k|g_j) \end{aligned} \tag{57.5}$$

Inverting the conditional probability $P(z_k|g_j)$ in Equation 57.5 with the application of Bayes' rule results in

$$P(r_i,g_j) = \sum_{k=1}^{K} P(z_k)P(r_i|z_k)P(g_j|z_k) \tag{57.6}$$

Following the likelihood principle, one determines $P(z_k)$, $P(r_i|z_k)$, and $P(g_j|z_k)$ by the maximization of the log-likelihood function

$$\mathscr{L} = \log P(R,G) = \sum_{i=1}^{M}\sum_{j=1}^{N} n(r_i,g_j)\log P(r_i,g_j) \tag{57.7}$$

where $n(r_i,g_j)$ denotes the number of the regions r_i that occurred in image g_j. From Equations 57.7 and 57.5 we derive that the model is a statistical mixture model (Mclachlan & Basford, 1988), which can be resolved by applying the EM technique (Dempster et al., 1977).

Model Fitting with EM

One powerful procedure for maximum likelihood estimation in hidden variable models is the EM method (Dempster et al., 1977). EM alternates in two steps iteratively: (i) an expectation (E) step where posterior probabilities are computed for the hidden variable z_k, based on the current estimates of the parameters, and (ii) a maximization (M) step, where parameters are updated to maximize the expectation of the complete-data likelihood $\log P(R,G,Z)$ for the given posterior probabilities computed in the previous E-step.

Applying Bayes' rule with Equation 57.5, we determine the posterior probability for z_k under (r_i,g_j):

$$P(z_k|r_i,g_j) = \frac{P(z_k)P(g_j|z_k)P(r_i|z_k)}{\sum_{k'=1}^{K} P(z_{k'})P(g_j|z_{k'})P(r_i|z_{k'})} \tag{57.8}$$

The expectation of the complete-data likelihood $\log P(R,G,Z)$ for the estimated $P(Z|R,G)$ derived from Equation 57.8 is

$$E\{\log P(R,G,Z)\} = \sum_{(i,j)=1}^{K}\sum_{i=1}^{M}\sum_{j=1}^{N} n(r_i,g_j)\log\left[P(z_{i,j})P(g_j|z_{i,j})P(r_i|z_{i,j})\right]P(Z|R,G) \tag{57.9}$$

where

$$P(Z|R,G) = \prod_{m=1}^{M}\prod_{n=1}^{N} P(z_{m,n}|r_m,g_n)$$

In Equation 57.9 the notation $z_{i,j}$ is the concept variable that is associated with the region-image pair (r_i,g_j). In other words, (r_i,g_j) belongs to concept z_t where $t=(i,j)$.

With the normalization constraint $\sum_{(i,j)=1}^{K} P(z_{i,j}|r_i,g_j) = 1$, Equation 57.9 further becomes:

$$E\{\log P(R,G,Z)\} = \sum_{l=1}^{K}\sum_{i=1}^{M}\sum_{j=1}^{N} n(r_i,g_j)\log[P(r_i|z_l)P(g_j|z_l)]P(z_l|r_i,g_j) + \\ + \sum_{l=1}^{K}\sum_{i=1}^{M}\sum_{j=1}^{N} n(r_i,g_j)\log[P(z_l)]P(z_l|r_i,g_j) \tag{57.10}$$

Maximizing Equation 57.10 with Lagrange multipliers to $P(z_l)$, $P(r_u|z_l)$, and $P(g_v|z_l)$, respectively, under the following normalization constraints

$$\sum_{k=1}^{K} P(z_k) = 1 \tag{57.11}$$

$$\sum_{k=1}^{K} P(z_k|r_i,g_j) = 1 \tag{57.12}$$

$$\sum_{i=1}^{M} P(r_i|z_l) = 1 \tag{57.13}$$

for any r_i, g_j, and z_l, the parameters are determined as

$$P(z_k) = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} n(r_i,g_j)P(z_k|r_i,g_j)}{\sum_{i=1}^{M}\sum_{j=1}^{N} u(r_i,g_j)} \tag{57.14}$$

$$P(r_u|z_l) = \frac{\sum_{j=1}^{N} n(r_u,g_j)P(z_l|r_u,g_j)}{\sum_{i=1}^{M}\sum_{j=1}^{N} u(r_i,g_j)P(z_l|r_i,g_j)} \tag{57.15}$$

$$P(g_v|z_l) = \frac{\sum_{i=1}^{M} n(r_i,g_v)P(z_l|r_i,g_v)}{\sum_{i=1}^{M}\sum_{j=1}^{N} u(r_i,g_j)P(z_l|r_i,g_j)} \tag{57.16}$$

Alternating Equation 57.8 with Equations 57.14–57.16 defines a convergent procedure that approaches a local maximum of the expectation in Equation 57.10. The initial values for $P(z_k)$, $P(g_j|z_k)$, and $P(r_i|z_k)$ are set to be the same as if the distributions of $P(Z)$, $P(G|Z)$, and $P(R|Z)$ are the uniform distributions; in other words, $P(z_k) = 1/K$, $P(r_i|z_k) = 1/M$, and $P(g_j|z_k) = 1/N$. We have found in the experiments that different initial values only affect the number of iterative steps to the convergence but have no effects on the converged values of them.

Estimating the Number of Concepts

The number of concepts, K, must be determined in advance to initiate the EM model fitting. Ideally, we would like to select the value of K that best represents the number of the semantic classes in the database. One readily available notion of the goodness of the fitting is the log-likelihood. Given this indicator, we apply the Minimum Description Length (MDL) principle (Rissanen, 1978, Rissanen, 1989) to select the best value of K. This can be operationalized as follows (Rissanen, 1989): choose K to maximize

$$\log(P(R,G)) - \frac{m_K}{2}\log(MN) \tag{57.17}$$

where the first term is expressed in Equation 57.7 and m_K is the number of the free parameters needed for a model with K mixture components. In the case of the proposed probabilistic model, we have

$$m_K = (K-1)+K(M-1)+K(N-1) = K(M+N-1)-1$$

As a consequence of this principle, when models using two values of K fit the data equally well, the simpler model is selected. In the database used in the experiments reported in Section 57.3.6, K is determined through maximizing Equation 57.17.

57.3.4 Posterior Probability Based Image Mining and Retrieval

Based on the probabilistic model, we can derive the posterior probability of each image in the database for every discovered concept by applying Bayes' rule as

$$P(z_k|g_j) = \frac{P(g_j|z_k)P(z_k)}{P(g_j)} \tag{57.18}$$

which can be determined using the estimations in Equations 57.14–57.16. The posterior probability vector $P(Z|g_j) = [P(z_1|g_j), P(z_2|g_j), \dots, P(z_K|g_j)]^T$ is used to quantitatively describe the semantic concepts associated with the image g_j. This vector can be treated as a representation of g_j (which originally has a representation in the M-dimensional "code word" space) in the K-dimensional *concept space* determined using the estimated $P(z_k|r_i,g_j)$ in Equation 57.8.

For each query image, after obtaining the corresponding "code words" as described in Section 57.3.2, we attain its representation in the discovered concept space by substituting it in the EM iteration derived in Section 57.3.3. The only difference is that $P(r_i|z_k)$ and $P(z_k)$ are fixed to be the values we have obtained for the whole database modeling (which are obtained in the indexing phase, i.e., to determine the concept space representation of every image in the database).

In designing a region-based image mining and retrieval methodology, there are two characteristics of the region representation that must be taken into consideration:

1. The number of the segmented regions in one image is normally small.
2. Not all regions in one image are semantically relevant to a given image; some are unrelated or even non-relevant; which regions are relevant or irrelevant depends on the user's querying subjectivity.

Incorporating the "code words" corresponding to unrelated or non-relevant regions would hurt the mining or retrieval accuracy because the occurrences of these regions in one image tend to "fool" the probabilistic model such that erroneous concept representations would be generated. To address the two characteristics in image mining and retrieval explicitly, we employ the relevance feedback for the similarity measurement in the concept space. Relevance feedback has been demonstrated as great potential to capture users' querying subjectivity both in text retrieval and in image retrieval (Vasconcelos & Lippman, 2000,Rui et al., 1997). Consequently, a mining and retrieval algorithm based on the relevance feedback strategy is designed to integrate the probabilistic model to deliver a more effective mining and retrieval performance.

In the algorithm, we move the query point in the "code word" token space toward the good example points (the relevant images labeled by the user) and away from the bad example points (the irrelevant images labeled by the user) such that the region representation has more supports to the probabilistic model. At the same time, the query point is expanded with the "code words" of the labeled relevant images. On the other hand, we construct a negative example "code word" vector by applying a similar vector moving strategy such that the constructed negative vector lies near the bad example points and away from the good example points. The vector moving strategy uses a form of Rocchio's formula (Rocchio, 1971). Rocchio's formula for relevance feedback and feature expansion has proven to be one of the best iterative optimization techniques in the field of information retrieval. It is frequently used to estimate the "optimal query" in relevance feedback for sets of relevant documents D_R and irrelevant documents D_I given by the user. The formula is

$$Q' = \alpha Q + \beta(\frac{1}{N_R} \sum_{j \in D_R} D_j) - \gamma(\frac{1}{N_I} \sum_{j \in D_I} D_j) \qquad (57.19)$$

where α, β, and γ are suitable constants; N_R and N_I are the number of documents in D_R and D_I, respectively; and Q' is the updated query of the previous query Q.

In the algorithm, based on the vector moving strategy and Rocchio's formula, in each iteration a modified query vector *pos* and a constructed negative example *neg* are computed; their representations in the discovered concept space are obtained and their similarities to each image in the database are measured through the cosine metric (Baeza-Yates & Ribeiro-Neto, 1999) of the corresponding vectors in the concept space, respectively. The retrieved images are ranked based on the similarity to *pos* as well as the dissimilarity to *neg*. The algorithm is described in Algorithm 3.

Algorithm 1: A semantic concept mining based retrieval algorithm

```
Input: $q$, "code word" vector of the query image
Output: Images retrieved for the query image $q$
Method:
Plug $q$ to the model to compute the vector $P(Z|q)$;
Retrieve and rank images based on the cosine similarity measure of the vectors $P(Z|q)$
    and $P(Z|g)$ of each image in the database;
$rs = \{rel_1, rel_2, \ldots, rel_a\}$, where $rel_i$ is a "code word" vector of each image labeled as
    relevant by the user on the retrieved result;
$is = \{ire_1, ire_2, \ldots, ire_b\}$, where $ire_j$ is a "code word" vector of each image labeled as
    irrelevant by the user on the retrieved result;
$pos = \alpha q + \beta(\frac{1}{a}\sum_{i=1}^{a} rel_i) - \gamma(\frac{1}{b}\sum_{j=1}^{b} ire_j)$;
$neg = \alpha(\frac{1}{b}\sum_{j=1}^{b} ire_j) - \gamma(\frac{1}{a}\sum_{i=1}^{a} rel_i)$;
for $k = 1$ to $K$ do
    Determine $P(z_k|pos)$ and $P(z_k|neg)$ with EM and Equation 57.18;
end for
$n = 1$;
while $n <= N$ do
    $sim1(g_n) = \frac{P(Z|pos) \bullet P(Z|g_n)}{\|P(Z|pos)\|\|P(Z|g_n)\|}$;
    $sim2(g_n) = \frac{P(Z|neg) \bullet P(Z|g_n)}{\|P(Z|neg)\|\|P(Z|g_n)\|}$;
    if $(sim1(g_n) > sim2(g_n))$ then
        $sim(g_n) = sim1(g_n) - sim2(g_n)$;
    else
        $sim(g_n) = 0$;
    end if
    Rank the images in the database based on $sim(g_n)$;
end while
```

We use the cosine metric to compute $sim1(\bullet)$ and $sim2(\bullet)$ in Algorithm 3 because the posterior probability vectors are the basis for the similarity measure in this proposed approach. The vectors are uniform, and the value of each component in the vectors is between 0 and 1. The cosine similarity is effective and ideal for measuring the similarity for the space composed of these kinds of vectors. The experiments reported in Section 57.3.6 show the effectiveness of the cosine similarity measure. At the same time, we note that Algorithm 3 itself is orthogonal to the selections of similarity measure metrics. The parameters α, β, and γ in Algorithm 3 are assigned a value of 1.0 in the current implementation of the prototype system for the sake of simplicity. However, other values may be used to emphasize the different weights between good sample points and bad sample points.

57.3.5 Approach Analysis

It is worth comparing the proposed probabilistic model and the fitting methodology with the existing region based statistical clustering methods in the image mining and retrieval literature, such as (Zhang & Zhang, 2004b, Chen et al., 2003). In the clustering methods, one typically associates a class variable with each image or each region in the database based on specific similarity metrics cast. One fundamental problem overlooked in such methods is that the semantic concepts of a region are typically not entirely determined by the features of the region itself; rather, they are dependent upon and affected by the contextual environment around the region in the image. In other words, a region in a different context in an image may convey a different concept. It is also noticeable that the degree of a specific region associated with several semantic concepts varies with different contextual region co-occurrences in an image. For example, it is likely that the *sand* “code word” conveys the concept of *beach* when it co-occurs in the context of the *water*, *sky*, and *people* “code words”; on the other hand, it becomes likely that the same *sand* “code word” conveys the concept of *African* with a high probability when it co-occurs in the context of the *plant* and *black* “code words”. Wang et al (Wang et al., 2001) attempted to alleviate the effect caused by this problem by using integrated region matching to incorporate similarity between two images for all their region pairs; this matching scheme, however, is heuristic such that it is impossible for a more rigorous analysis.

The probabilistic model we have described addresses these problems quantitatively and analytically in an optimal framework. Given a region in an image the conditional probability of each concept and the conditional probability of each image in a concept are iteratively determined to fit the model representing the database as formulated in Equations 57.8 and 57.16. Since the EM technique always converges to a local optimality, from the experiments reported in Section 57.3.6, we have found that the local optimum is satisfactory for typical image data mining and retrieval applications. The effectiveness of this methodology in real image databases is demonstrated in the experimental analysis presented in Section 57.3.6. To find the global maximum is computationally intractable for a large-scale database, and the advantage of such model fitting compared to the model fitting obtained through this proposed approach is not obvious and is under further investigation.

With the proposed probabilistic model, we are able to concurrently obtain $P(z_k|r_i)$ and $P(z_k|g_j)$ such that both regions and images have an interpretation in the concept space simultaneously, while typical image clustering based approaches, such as (Jing et al., 2004), do not have this flexibility. Since in the proposed scheme, every region and/or image may be represented as a weighted sum of the components along the discovered concept axes, the proposed model acts as a factoring analysis (Mclachlan & Basford, 1988), yet the same model offers important advantages, such as that each weight has a clear probabilistic meaning and the factoring is two-fold, i.e., both regions and images in the database have probabilistic representations with the discovered concepts.

Another advantage of the proposed methodology is its capability to reduce the dimensionality. The image similarity comparison is performed in a derived K-dimensional concept space Z instead of in the original M-dimensional "code word" token space R. Note that typically $K << M$, as has been demonstrated in the experiments reported in Section 57.3.6. The derived subspace represents the hidden semantic concepts conveyed by the regions and the images, while the noise and all the non-intrinsic information are discarded in the dimensionality reduction, which makes the semantic comparison of regions and images more effective and efficient. The coordinates in the concept space for each image as well as for each region are determined by automatic model fitting. The computation requirement in the lower-dimensional concept space is reduced as compared with that required in the original "code word" space.

Algorithm 3 integrates the posterior probability of the discovered concepts with the query expansion and the query vector moving strategy in the "code word" token space. Consequently, the accuracy of the representation of the semantic concepts of a user's query is enhanced in the "code word" token space, which also improves the accuracy of the position obtained for the query image in the concept space. Moreover, the constructed negative example *neg* improves the discriminative power of the probabilistic model. Both the similarity to the modified query representation and the dissimilarity to the constructed negative example in the concept space are employed.

57.3.6 Experimental Results

We have implemented the approach in a prototype system on a platform of a Pentium IV 2.0 GHz CPU and 256 MB memory. The interface of the system is shown in Figure 57.13. The following reported evaluations are performed on a general-purpose color image database containing 10,000 images from the COREL collection with 96 semantic categories. Each semantic category consists of 85–120 images. In Table 57.1, exemplar categories in the database are provided. We note that the category information in the COREL collection is only used to ground-truth the evaluation, and we do not make use of this information in the indexing, mining, and retrieval procedures. Figure 57.7 shows a few examples of the images in the database.

To evaluate the image retrieval performance, 1,500 images are randomly selected from all the categories as the query set. The relevancy of the retrieved images is subjectively examined by users. The ground truth used in the mining and retrieval experiments is the COREL category label if the query image is in the database. If the query image is a new image outside the database, users' specified relevant images in the mining and retrieval results are used to calculate the mining and retrieval accuracy statistics. Unless otherwise noted, the default results of the experiments are the averages of the top 30 returned images for each of the 1,500 queries.

In the experiments, the parameters of the image segmentation algorithm (Wang et al., 2001) are adjusted with the consideration of the balance of the depiction detail and the computation complexity such that there is an average of 8.3207 regions in each image. To determine the size of the visual token catalog, different numbers of the "code words" are selected and evaluated. The average precisions (without the query expansion and movement) within the top 20, 30, and 50 images, denoted as P(20), P(30), and P(50), respectively, are shown in Figure 57.8. It indicates that the general trend is that the larger the visual token catalog size, the higher the mining and retrieval accuracy. However, a larger visual token catalog size means a larger number of image feature vectors, which implies a higher computation complexity in the process of the hidden semantic concept discovery. Also, a larger visual token catalog leads to a larger storage space. Therefore, we use 800 as the number of the "code words", which

Table 57.1. Examples of the 96 categories and their descriptions. Reprint from (Zhang & Zhang, 2007) ©2007 IEEE Signal Processing Society Press.

ID	Category description
1	reptile, animal, rock
2	Britain, royal events, queen, prince, princess
3	Africa, people, landscape, animal
4	European, historical building, church
5	woman, fashion, model, face, cloth
6	hawk, sky
7	New York City, skyscrapers, skyline
8	mountain, landscape
9	antique, craft
10	Easter egg, decoration, indoor, man-made
11	waterfall, river, outdoor
12	poker cards
13	beach, vacation, sea shore, people
14	castle, grass, sky
15	cuisine, food, indoor
16	architecture, building, historical building
..	......

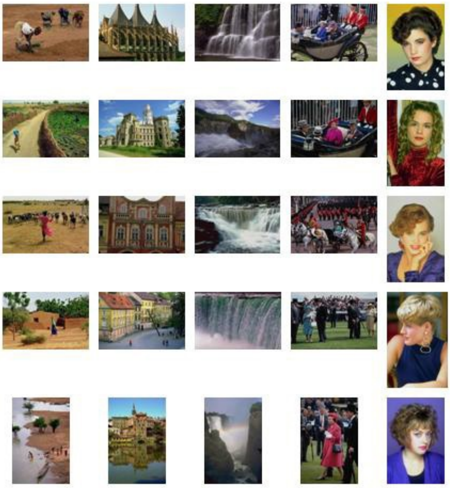


Fig. 57.7. Sample images in the database. The images in each column are assigned to one category. From left to right, the categories are Africa rural area, historical building, waterfalls, British royal event, and model portrait, respectively.

corresponds to the first turning point in Figure 57.8. Since there are a total of 83,307 regions in the database, on average each “code word” represents 104.13 regions.

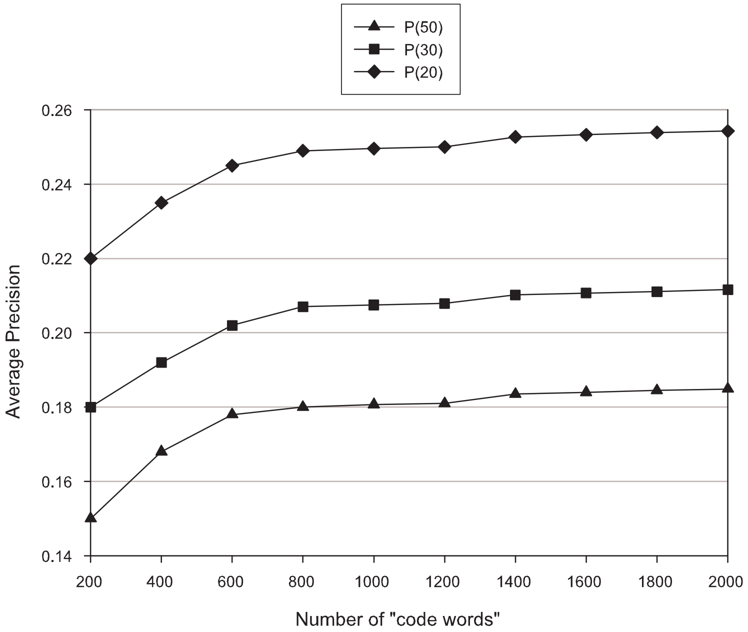


Fig. 57.8. Average precision (without the query expansion and movement) for different sizes of the visual token catalog. Reprint from (Zhang & Zhang, 2007) ©2007 IEEE Signal Processing Society Press and from (Zhang & Zhang, 2004a) ©2004 IEEE Computer Society Press.

Applying the method of estimating the number of the hidden concepts described in Section 57.3.3, the number of the concepts is determined to be 132. Performing the EM model fitting, we have obtained the conditional probability of each “code word” to every concept, i.e., $P(r_i|z_k)$. Manual examination of the visual content of the region sets corresponding to the top 10 highest “code words” in every semantic concept reveals that these discovered concepts indicate semantic interpretations, such as “people”, “building”, “outdoor scenery”, “plant”, and “automotive race”. Figure 57.9 shows several exemplar concepts discovered and the top regions corresponding to $P(r_i|z_k)$ obtained.

In terms of the computational complexity, despite the iterative nature of EM, the computing time for the model fitting at $K = 132$ is acceptable (less than 1 second). The average number of iterations upon convergence for one image is less than 5.

We give an example for discussion. Figure 57.10 shows one image, *Im*, belonging to the “medieval building” category in the database. *Im* (i.e., Figure 57.10(a)) has 6 “code words” associated. Each “code word” is presented using a unique color graphically in Figure 57.10(b). For the sake of discussion, the indices for these “code words” are assigned to be 1–6, respectively.

Figure 57.11 shows the $P(z_k|r_i,Im)$ for each “code word” r_i (represented as a different color) and the posterior probability $P(z_k|Im)$ after the first iteration and the last iteration in the

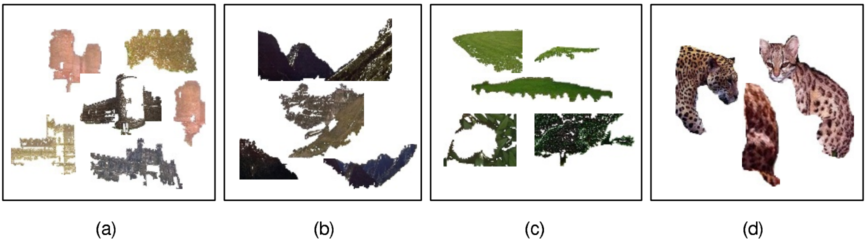


Fig. 57.9. The regions with the top $P(r_i|z_k)$ to the different concepts discovered. (a) "castle"; (b) "mountain"; (c) "meadow and plant"; (d) "cat". Reprint from (Zhang & Zhang, 2007) ©2007 IEEE Signal Processing Society Press.

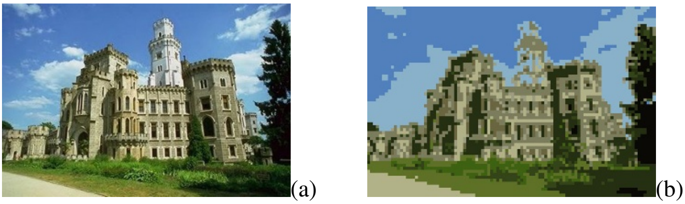


Fig. 57.10. Illustration of one query image in the "code word" space. (a) Image *Im*; (b) "code word" representation. Reprint from (Zhang & Zhang, 2007) ©2007 IEEE Signal Processing Society Press.

course of the EM model fitting. Here the 4 concepts with highest $P(z_k|Im)$ are shown. From left to right in Figure 57.11, they represent "plant", "castle", "cat", and "mountain", respectively, interpreted through manual examination. As is seen in the figure, the "castle" concept has indeed the highest weight after the first iteration; nevertheless, the other three concepts still account for more than half of the probability. The probability distribution changes after several EM iterations, since the proposed probabilistic model incorporates co-occurrence patterns between the "code words"; i.e., $P(z_k|r_i)$ is not only related to one "code word" (r_i) but is also related to all the co-occurring "code words" in the image. For example, although "code word" 2, which accounts for "meadow", has higher fitness in the concept "plant" after the first iteration, the context of the other regions in image *Im* increases the probability that this "code word" is related to the concept "castle" and decreases its probability related to "plant" as well.

Figure 57.12 shows the similar plot to Figure 57.11 except that we apply the relevance feedback based query expansion and moving strategy to image *Im* as described in the Algorithm 3. The "code word" vector of image *Im* is expanded to contain 10 "code words". Compared with Figure 57.11, it is clear that with the expansion of the relevant "code words" to *Im* and the query moving strategy toward the relevant image set, the posterior probabilities favoring the concept "castle" increase while the posterior probabilities favoring other concepts decrease substantially, resulting in an improved mining and retrieval precision, accordingly.

To show the effectiveness of the probabilistic model in image mining and retrieval, we have compared the accuracy of this methodology with that of UFM (Chen & Wang, 2002) proposed by Chen and Wang. UFM is a method based on the fuzzified region representation to build region-to-region similarity measures for image retrieval; it is an improvement of their early work SIMPLIcity (Wang et al., 2001). The reasons why we compare this proposed approach with UFM are: (1) the UFM system is available to us; and (2) UFM reflects the

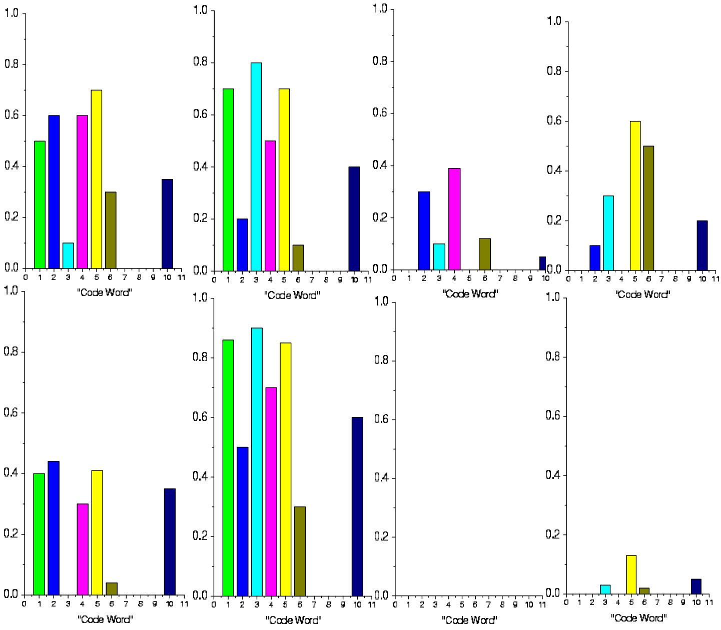


Fig. 57.11. $P(z_k|r_i,Im)$ (each color column for a "code word") and $P(z_k|Im)$ (rightmost column in each bar plot) for image Im for the four concept classes (semantically related to "plant", "castle", "cat", and "mountain", from left to right, respectively) after the first iteration (first row) and the last iteration (second row). Reprint from (Zhang & Zhang, 2007) ©2007 IEEE Signal Processing Society Press.

performance of the state-of-the-art image mining and retrieval performance. In addition, the same image segmentation and feature extraction methods are used in UFM such that a fair comparison on the performance between the two systems is ensured. Figure 57.13 shows the top 16 retrieved images by the prototype system and as well as by UFM, respectively, using image Im as a query.

More systematic comparison results on the 1,500 query image set are reported in Figure 57.14. Two versions of the prototype (one with the query expansion and moving strategy and the other without) and UFM are evaluated. It is demonstrated that the performances of the probabilistic model in both versions of the prototype have higher overall precisions than that of UFM, and the query expansion and moving strategy with the interaction of the constructed negative examples boost the mining and retrieval accuracy significantly.

57.4 Summary

In this chapter we have introduced the new, emerging area called multimedia data mining. We have given a working definition of what this area is about; we have corrected a few misconceptions that typically exist in the related research communities; and we have given a typical

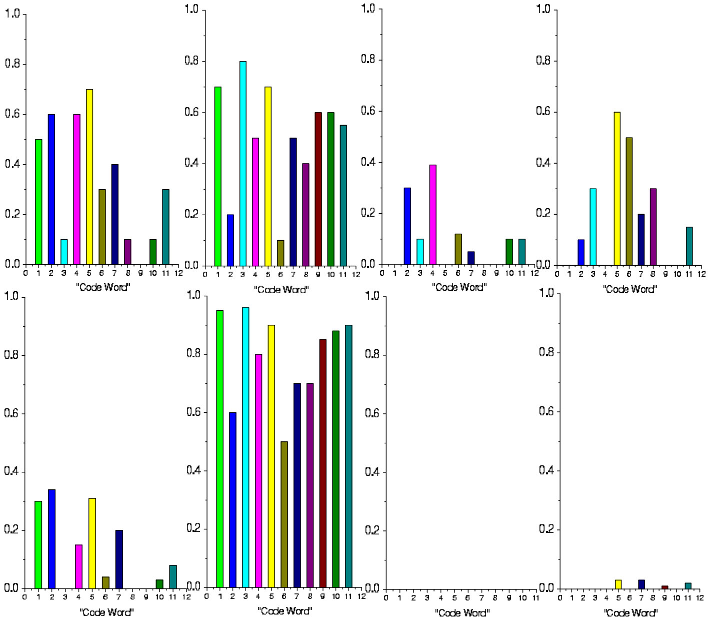


Fig. 57.12. The similar plot to Figure 57.11 with the application of the query expansion and moving strategy. Reprint from (Zhang & Zhang, 2007) ©2007 IEEE Signal Processing Society Press.

architecture for a multimedia data mining sytem or methodology. Finally, in order to showcase what a typical multimedia data mining system does and how it works, we have given an example of a specific method for semantic concept discovery in an imagery database.

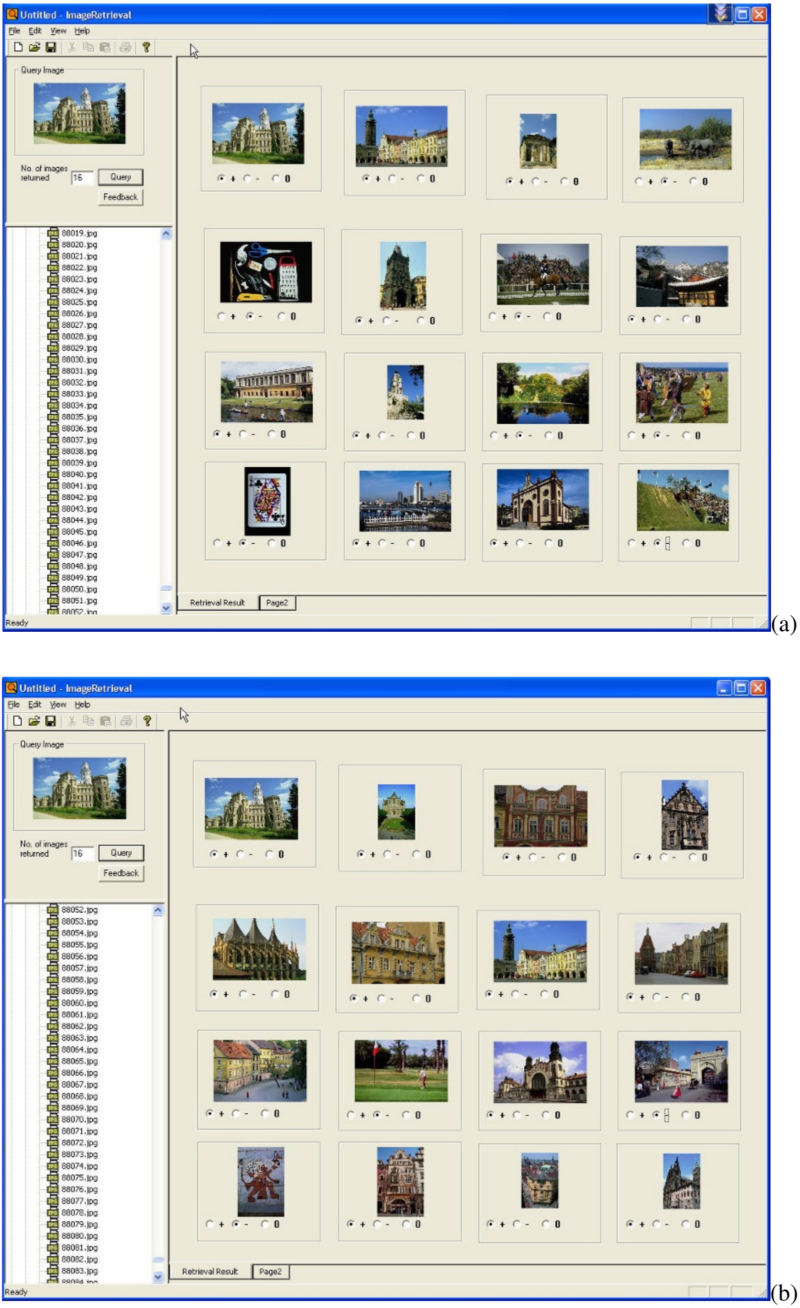
Multimedia data mining, though it is a new and emerging area, has undergone an independent and rapid development over the last few years. A systematic introduction to this area may be found in (Zhang & Zhang, 2008) as well as the further readings contained in the book.

Ackonwledgments

This work is supported in part by the National Science Foundation through grants IIS-0535162 and IIS-0812114. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Baeza-Yates, R. & Ribeiro-Neto, B. (1999). Modern Information Retrieval. Addison-Wesley.



No. of images returned 16

Fig. 57.13. Retrieval performance comparisons between UFM and the prototype system using image *Im* in Figure 57.10 as the query. (a) Images returned by UFM (9 of the 16 images are relevant). (b) Images returned by the prototype system (14 of the 16 images are relevant).

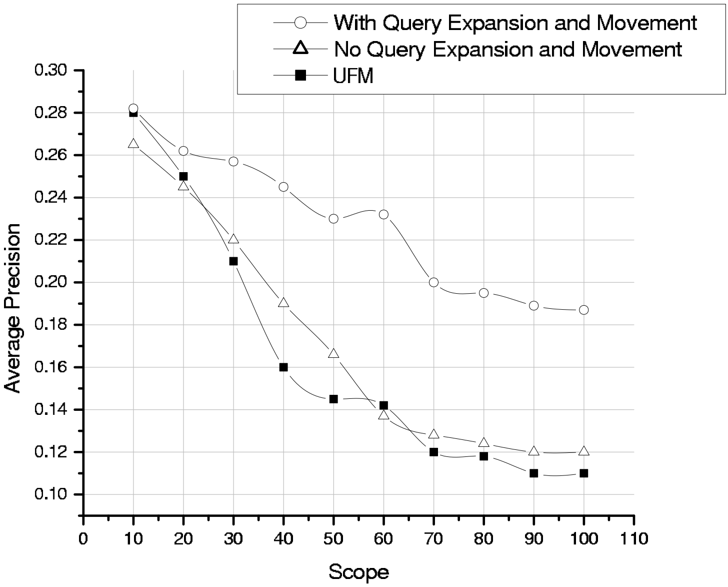


Fig. 57.14. Average precision comparisons between the two versions of the prototype and UFM. Reprint from (Zhang & Zhang, 2007) ©2007 IEEE Signal Processing Society Press and from (Zhang & Zhang, 2004a) ©2004 IEEE Computer Society Press.

Barnard, K., Duygulu, P., d.Freitas, N., Blei, D. & Jordan, M. I. (2003). Journal of Machine Learning Research 3, 1107–1135.

Barnard, K. & Forsyth, D. (2001). In The International Conference on Computer Vision vol. II, pp. 408–415,.

Blei, D., Ng, A. & Jordan, M. (2001). In The International Conference on Neural Information Processing Systems.

Carbonetto, P., d. Freitas, N. & Barnard, K. (2004). In The 8th European Conference on Computer Vision.

Carbonetto, P., d. Freitas, N., Gustafson, P. & Thompson, N. (2003). In The 9th International Workshop on Artificial Intelligence and Statistics.

Carson, C., Belongie, S., Greenspan, H. & Malik, J. (2002). IEEE Trans. on PAMI 24, 1026–1038.

Castleman, K. (1996). Digital Image Processing. Prentice Hall, Upper Saddle River, NJ.

Chen, Y. & Wang, J. (2002). IEEE Trans. on PAMI 24, 1252–1267.

Chen, Y., Wang, J. & Krovetz, R. (2003). In the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval pp. 193–200,, Berkeley, CA.

Dempster, A., Laird, N. & Rubin, D. (1977). Journal of the Royal Statistical Society, Series B 39, 1C38.

Duygulu, P., Barnard, K., d. Freitas, J. F. G. & Forsyth, D. A. (2002). In The 7th European Conference on Computer Vision vol. IV, pp. 97–112,, Copenhagon, Denmark.

Faloutsos, C. (1996). Searching Multimedia Databases by Content. Kluwer Academic Publishers.

Faloutsos, C., Barber, R., Flickner, M., Hafner, J., Niblack, W., Petkovic, D. & Equitz, W. (1994). Journal of Intelligent Information Systems 3, 231–262.

Feng, S. L., Manmatha, R. & Lavrenko, V. (June, 2004). In The International Conference on Computer Vision and Pattern Recognition, Washington, DC.

Flickner, M., Sawhney, H., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D. & Yanker, P. (1995). IEEE Computer 28, 23–32.

Furht, B., ed. (1996). Multimedia Systems and Techniques. Kluwer Academic Publishers.

Greenspan, H., Dvir, G. & Rubner, Y. (2004). Journal of Computer Vision and Image Understanding 93, 86–109.

Greenspan, H., Goldberger, J. & Ridel, L. (2001). Journal of Computer Vision and Image Understanding 84, 384–406.

Han, J. & Kamber, M. (2006). Data Mining — Concepts and Techniques. 2 edition, Morgan Kaufmann.

Hofmann, T. (2001). Machine Learning 42, 177C196.

Hofmann, T. & Puzicha, J. (1998). AI Memo 1625.

Hofmann, T., Puzicha, J. & Jordan, M. I. (1996). In The International Conference on Neural Information Processing Systems.

Huang, J. & et al., S. R. K. (1997). In IEEE Int'l Conf. Computer Vision and Pattern Recognition Proceedings, Puerto Rico.

Jain, R. (1996). In Multimedia Systems and Techniques, (Furht, B., ed.),. Kluwer Academic Publishers.

Jeon, J., Lavrenko, V. & Manmatha, R. (2003). In the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Jing, F., Li, M., Zhang, H.-J. & Zhang, B. (2004). IEEE Trans. on Image Processing 13.

Kohonen, T. (2001). Self-Organizing Maps. Springer, Berlin, Germany.

Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V. & Saarela, A. (2000). IEEE Trans. on Neural Networks 11, 1025–1048.

Ma, W. & Manjunath, B. S. (1995). In Internation Conference on Image Processing pp. 2256–2259,.

Ma, W. Y. & Manjunath, B. (1997). In IEEE Int'l Conf. on Image Processing Proceedings pp. 568–571,, Santa Barbara, CA.

Maimon O., and Rokach, L. Data Mining by Attribute Decomposition with semiconductors manufacturing case study, in Data Mining for Design and Manufacturing: Methods and Applications, D. Braha (ed.), Kluwer Academic Publishers, pp. 311–336, 2001.

Manjunath, B. S. & Ma, W. Y. (1996). IEEE Trans. on Pattern Analysis and Machine Intelligence 18.

Mclachlan, G. & Basford, K. E. (1988). Mixture Models. Marcel Dekker, Inc., Basel, NY.

Moghaddam, B., Tian, Q. & Huang, T. (2001). In The International Conference on Multimedia and Expo 2001.

Pentland, A., Picard, R. W. & Sclaroff, S. (1994). In SPIE-94 Proceedings pp. 34–47,.

Rissanen, J. (1978). Automatica 14, 465–471.

Rissanen, J. (1989). Stochastic Complexity in Statistical Inquiry. World Scientific.

Rocchio, J. J. J. (1971). In The SMART Retreival System — Experiments in Automatic Document Processing pp. 313–323. Prentice Hall, Inc Englewood Cliffs, NJ.

Rokach L., Mining manufacturing data using genetic algorithm-based feature set decomposition, Int. J. Intelligent Systems Technologies and Applications, 4(1):57-78, 2008.

Rokach, L. and Maimon, O. and Averbuch, M., Information Retrieval System for Medical Narrative Reports, Lecture Notes in Artificial intelligence 3055, page 217-228 Springer-Verlag, 2004.

Rui, Y., Huang, T. S., Mehrotra, S. & Ortega, M. (1997). In IEEE Workshop on Content-based Access of Image and Video Libraries, in conjunction with CVPR'97 pp. 82–89,.

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A. & Jain, R. (2000). IEEE Trans. on Pattern Analysis and Machine Intelligence 22, 1349–1380.

Steinmetz, R. & Nahrstedt, K. (2002). Multimedia Fundamentals — Media Coding and Content Processing. Prentice-Hall PTR.

Subrahmanian, V. (1998). Principles of Multimedia Database Systems. Morgan Kaufmann.

Vasconcelos, N. & Lippman, A. (2000). In IEEE Workshop on Content-based Access of Image and Video Libraries (CBAIVL'00), Hilton Head, South Carolina.

Wang, J., Li, J. & Wiederhold, G. (2001). IEEE Trans. on PAMI 23.

Wood, M. E. J., Campbell, N. W. & Thomas, B. T. (1998). In ACM Multimedia 98 Proceedings, Bristol, UK.

Zhang, R. & Zhang, Z. (2004a). In IEEE International Conference on Computer Vision and Pattern Recogntion (CVPR) 2004, Washington, DC.

Zhang, R. & Zhang, Z. (2004b). EURASIP Journal on Applied Signal Processing 2004, 871–885.

Zhang, R. & Zhang, Z. (2007). IEEE Transactions on Image Processing 16, 562–572.

Zhang, Z. & Zhang, R. (2008). Multimedia Data Mining — A Systematic Introduction to Concepts and Theory. Taylor & Francis.

Zhou, X. S., Rui, Y. & Huang, T. S. (1999). In IEEE Conf. on Image Processing Proceedings.

Zhu, L., Rao, A. & Zhang, A. (2002). ACM Transaction on Information Systems 20, 224–257.