CHAPTER 1

Introduction

With the explosion of information brought about by this Multimedia Age, the question of how such information might be effectively harvested, archived, and analysed, remains a monumental challenge facing today's research community. The processing of such information, however, is often fraught with the need for conceptual interpretation—a relatively simple task for humans, yet arduous for computers. In order to handle the oppressive volumes of information that are becoming readily accessible in consumer and industrial sectors, some level of automation is desirable.

Automation requires computational systems that exhibit some degree of intelligence, in terms of the ability of a system to formulate its own models of the data in question with little or no user intervention. Such systems must be able to make basic decisions about what information is actually important and what is not. In effect, like a human user, the system must be able to discover characteristic properties of the data in some appropriate manner, without a teacher. This process is known as *unsupervised learning* (sometimes referred to as *clustering* or *unsupervised pattern classification*; an essentially pure form of data mining).

This book primarily introduces a new approach to the general problem of unsupervised learning, based on the principles of *dynamic self-organization*. Inspired by the relative success of other popular research on self-organizing neural networks for data clustering and feature extraction, this book presents new members within the family of generative, Self-Organizing Maps, namely: the self-organizing tree map (SOTM) and its advanced form, the *self-organizing hierarchical variance map* (SOHVM). While the devised approach is essentially generic, the core application considered in this book is the automatic, unsupervised data clustering for multimedia applications and unsupervised segmentation of microbiological image data.

1.1 PART I: THE SELF-ORGANIZING METHOD

Computational technologies based on Artificial Neural Networks (ANN) have been the focus of much research into the problem of unsupervised learning, in particular,

Unsupervised Learning: A Dynamic Approach, First Edition.
Matthew Kyan, Paisarn Muneesawang, Kambiz Jarrah, and Ling Guan.

for network architectures that are based on principles of Self-Organization. Such principles are in many ways centered on Turing's initial observation in 1952 [1], namely, that Global order can arise from Local interactions. With much support from neurobiological research, such mechanisms are believed to be analogous to the organization that takes place in the human brain.

Clustering algorithms use unsupervised learning rules to group unlabeled training data into similar or dense clusters. Unsupervised training algorithms depend upon internally generated error measures, which are derived solely from training data. The network has no knowledge of the correct answer during training and, consequently, must derive the errors and the necessary weight modifications directly from the statistics of the training data. As a result, input patterns are stored as a set of cluster prototypes or exemplars—representations or natural groupings of similar data. In forming a description of an unknown set of data, such network architectures are characterized by their adherence to four key properties [2]: synaptic self-amplification for mining correlated stimuli, competition over limited resources, cooperative encoding of information, and the implicit ability to encode pattern redundancy as knowledge. Such principles are, in many ways, a reflection of Turing's observations previously discussed.

Part I of this book consists of Chapters 2 and 3. It gives an extensive review of the general problems of unsupervised clustering, with emphasis placed on the inherent relationship that exists between unsupervised learning and Self-Organization. The unsupervised learning problem is first defined with respect to the concepts of similarity and distance. A survey of unsupervised techniques from the broader field is then conducted to establish the context for more focused surveys on self-organization-based principles and architectures. The issue of validating unsupervised clustering solutions in the absence of a ground truth is also addressed.

1.2 PART II: DYNAMIC SELF-ORGANIZATION FOR IMAGE FILTERING AND MULTIMEDIA RETRIEVAL

Multimedia processing has seen impressive growth in the past decade in terms of both theoretical development and applications. It represents a leading technology in a number of important areas that warrant significant need for data mining, namely, digital telecommunications, multimedia systems, high dimensional image analysis and visualization, information retrieval, biology, robotics and manufacturing, and intelligent sensing systems. Inherently unsupervised in nature, neural network architectures based on principles of Self-Organization appear to be a natural fit.

In Part II of this book, the SOTM and its recently successful application in multimedia processing is presented. This neural network architecture incorporates hierarchical properties by virtue of its growth, in a manner that is flexible in terms of revealing the underlying data space without being constrained by an imposed topological framework. As such, the SOTM exhibits many desirable properties over traditional self-organizing feature map (SOFM) based strategies. Chapter 4 of the

book will provide an in-depth coverage of this architecture. Chapters 5 and 6 will then cover a series of pertinent real-world applications with regard to the processing of multimedia data. This includes problems in image-processing techniques, such as the automated modeling and removal of impulse noise in digital images, and problems in image classification in multimedia indexing and retrieval.

In Chapter 4, the SOTM algorithm is explored and developed, wherein a number of enhancements and modifications are proposed, justified, and tested, with the goal of rendering the SOTM more robust under application to different datasets. Specifically, alternative modalities for hierarchical control and learning are considered, in addition to more appropriate stopping criteria linked to aspects of the input data. The SOTM is then explored as a means of segmenting biofilm images, where its strengths and flexibility as a dynamic clustering model for segmentation are explored. Limitations and deficiencies of the SOTM are also identified.

In Chapter 5, the SOTM is applied to the automated modeling and removal of impulse noise in digital images. Improving the quality of images degraded by noise is a classic problem in image processing [3]. In the early stages of signal and image processing, linear filters were the primary tools for noise cleaning. Later, the development of nonlinear filtering techniques for signal and image processing was spurred by some drawbacks of linear filters [4]. However, one problem with nonlinear filers such as the median filter is that they remove the fine details in the image and change the signal structure. In addition, improved nonlinear filters, such as the weighted median filter, multistage median filter, and nonlinear mean filters, have better detail-preserving characteristics at the expense of poorer noise suppression. Here, a novel approach for suppressing impulse noise in digital images is proposed for effectively preserving more image detail than previously proposed methods. The noise removal system, shown in Figure 1.1a, consists of two steps: the detection of the noise and the reconstruction of the image. As the SOTM network has the capability to classify pixels in an image, it is employed to detect the impulses. A noise-exclusive median (NEM) filtering algorithm and a noise-exclusive arithmetic mean (NEAM) filtering algorithm are proposed to restore the image. This system is able to detect noise locations accurately, and thus, achieves the best possible restoration of images corrupted by impulse noise.

In Chapter 6, the SOTM is applied to problems in image classification in multimedia indexing and retrieval. The system architecture is shown in Figure 1.1b. In multimedia database retrieval, relevance feedback (RF) is a popular and effective way to improve the performance of image re-ranking and retrieval. However, RF needs a high level of human participation, which often leads to excessive subjective errors. Here, an automatic RF is present, using the SOTM, which minimizes user participation, providing a more user-friendly environment and avoiding errors caused by excessive human involvement. Unlike the conventional retrieval system, where the user's direct input is required in the execution of the RF algorithm, SOTM estimation is now adopted to guide the adaptation of the RF parameters. As shown in Figure 1.1b, the initially retrieved samples are labeled with the unsupervised module, and image re-ranking is performed by the pseudo-labeled samples. As a result, instead of

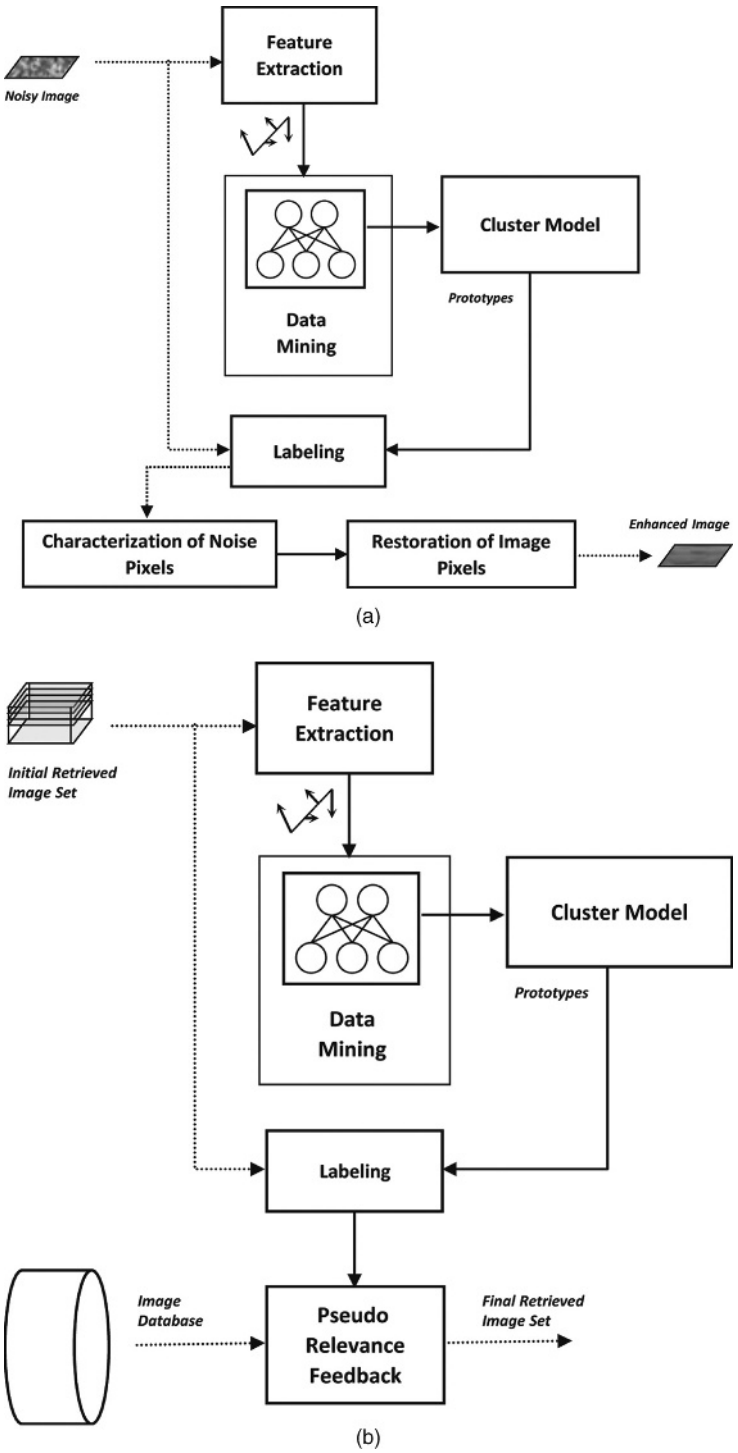

FIGURE 1.1 Unsupervised Learning–based framework for (a) automated modeling and removal of impulse noise in digital images and (b) image classification in multimedia indexing and retrieval.

imposing a greater responsibility on the user, independent learning can be integrated to improve retrieval accuracy. This makes it possible to obtain either a fully automatic or a semiautomatic RF system suitable for practical applications.

1.3 PART III: DYNAMIC SELF-ORGANIZATION FOR IMAGE SEGMENTATION AND VISUALIZATION

Much emphasis of this book is placed on Part III, on the developments of the SOHVM and its application in the unsupervised segmentation and visualization of microbiological image data. With recent advances in imaging, computer, and optical modalities for microscopy, a paradigm shift away from the purely observational toward the extraction of more quantitative information seems to be taking place. As such, data mining techniques are thought to serve as a useful basis for further processing stages. To this end this book demonstrates the capability of the newly proposed SOHVM over its predecessors and other popular self-organizing and partition-based clustering algorithms, for formulating a relatively stable clustering solution. Furthermore, avenues are explored for how the model can extract and use data associations discovered between clusters.

In the interests of assisting biologists in exploring previously unseen, unlabeled image data, unsupervised methods for attaining useful data-driven segmentations are explored, serving as a useful basis for further processing stages such as visualization or quantitative analysis.

In general, the approach taken is to identify characteristic biological materials present in the data, by identifying natural groupings (clusters) of similar voxel patterns—where an individual voxel pattern may be thought of as a vector of one or more different attributes, to which we refer as features. The framework for the approach taken is summarized in Figure 1.2.

In the most basic example, a voxel pattern from a single channel image might comprise a single feature only, namely, its intensity value. Alternatively, over a three-channel image, a voxel pattern may be a three-tuple vector, with one dimension for each channel. Under this framework, higher level, n-dimensional pattern vectors also become possible, allowing for the possibility of fusing local regional or other information extracted from the image into the description of each voxel. For instance, at a later stage in this study, the use of a classic texture feature (from the signal/image processing community) is incorporated into the description of a voxel for a single channel dataset, effectively transforming it into an 18+ tuple pattern vector.

Armed with an n-dimensional input data/feature space of actual pattern vectors, unsupervised learning or data clustering algorithms are aimed at parsing the space of possible patterns so as to locate regions of density that characterize the underlying data distribution. In this way, specific data samples may be associated and conglomerated into such characteristic groups. Each grouping is such that the data within demonstrates a certain level of *homogeneity*, with respect to some predefined similarity metric. Such metrics are typically implemented in the form of a *distance* mechanism, where the distance between two sample patterns gives a quantitative

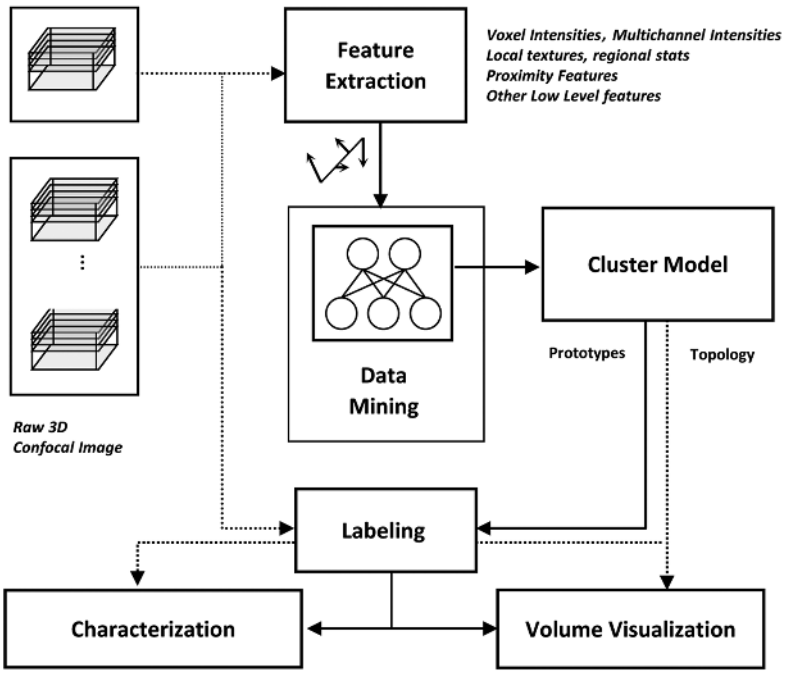

FIGURE 1.2 Unsupervised Learning–based framework for mining segmentations for visualization and characterization of microbiological image data.

measure of their relative dissimilarity (i.e., the smaller the distance, the more similar the patterns).

While there exist many classes of unsupervised learning algorithms in the literature (reviewed in Chapter 2), relatively few offer solutions to the problem of finding an appropriate number of pattern groups in which to break up the data at runtime. In fact, relatively few algorithms exist that are able to dynamically construct an appropriate number of groups as the data is parsed. The majority of unsupervised techniques available today are fixed, in that the number of groups (K) in which to partition the data is an argument to the algorithm, and it is more or less enforced. The current state of the art then, is to run a fixed algorithm multiple times, each with a different value for K, and then to evaluate the most suitable solution in an a posteriori manner, based on cluster validity analysis [5].

If an algorithm is able to dynamically construct a description of the input space throughout the course of parsing the data, then it seems natural that such a mechanism may be better suited to estimating an appropriate number of groups *on the fly*. ANN modeled on principles of dynamic (or generative) Self-Organization seem a natural fit toward this goal, as they possess the unique ability to grow and adapt over the course of time, while demonstrating heightened ability to generalize across a wide range of

linear and nonlinear input pattern stimuli, by virtue of their inherently associative and parallel nature. Exhibiting properties found to have neurobiological support [6], self-organizing-based networked systems reflect a realization of what we know to be at work in the cerebral cortex in the processing of information, namely, the existence of topologically ordered mappings of sensory inputs such as tactile, visual, and acoustic stimuli [2]. Inspired by this notion, the proposal of a new algorithm for unsupervised learning, based on dynamic Self-Organization, forms the core of this book.

In Chapter 7, a new model for unsupervised learning based on dynamic Self-Organization is proposed, namely, the SOHVM. The impetus and motivation for the new model arise out of the desire to overcome limitations identified with the SOTM. In addition, the new model embeds a mechanism to adaptively extract higher level associations from the data (interclass or topological relationships), as well as a mechanism for estimating an appropriate number of optimal classes at runtime. The performance and characteristics of the new model are then demonstrated with both the SOTM and other algorithms through a series of visual simulations.

In Chapter 8, the first half of the chapter presents a more concrete validation of both the modified SOTM and newly proposed SOHVM model. Specifically, an extensive cluster validity analysis is performed, drawing comparisons across a range of popular unsupervised clustering models. Issues of regularity, quality, and optimality are addressed with respect to synthetic and real-world data. The second half of the chapter then returns to the problem of segmenting microbiological image data, and in this regard, the performance of both the SOHVM and modified SOTM are demonstrated against other techniques. The discussion of the experiments on microbiological data then concludes with an example for how topological information mined by the SOHVM can be used to simplify the three-dimensional (3D) visualization of a large stack of chromosome data.

1.4 FUTURE DIRECTIONS

The focal points of the book lay in the design and development of two models for unsupervised leaning or data clustering, based on dynamic self-organization—namely, the self-organizing tree map (SOTM) and the self-organizing hierarchical variance map (SOHVM). Specific applications presented outline the utility of these models in applications including the automated modeling and removal of impulse noise in digital images; image classification in multimedia indexing and retrieval; and segmentation and visualization of biomedical image data.

The real advantage of creating a self-organizing clustering lies in the functionality of the resulting topological map. Mining the topology can be leveraged for very specific tasks. The major categories of tasks are

- Dynamic navigation through information repositories, applied for image re-ranking and visualization, video browsing, summarization, and retrieval.

- Interactive knowledge-assisted visualization, applied for volume exploration of volumetric multimedia datasets through direct volume rendering techniques to actively re-render the scene based on the user's current cluster of interest.
- Temporal data analysis using trajectories, applied for video shot boundary detection and gesture recognition.

In Chapter 9, the book crystallizes the main findings of this work and offers some recommendations of avenues for future research.