Chapter IX
Pattern Mining and Clustering on Image Databases

Marinette Bouet
LIMOS, Blaise Pascal University-Clermont-Ferrand, France

Pierre Gançarski
LSIIT-AFD-Louis Pasteur University, France

Marie-Aude Aufaure
Supélec—INRIA, France

Omar Boussaïd
University LUMIERE Lyon, France

ABSTRACT

Analysing and mining image data to derive potentially useful information is a very challenging task. Image mining concerns the extraction of implicit knowledge, image data relationships, associations between image data and other data or patterns not explicitly stored in the images. Another crucial task is to organise the large image volumes to extract relevant information. In fact, decision support systems are evolving to store and analyse these complex data. This chapter presents a survey of the relevant research related to image data processing. We present data warehouse advances that organise large volumes of data linked with images, and then we focus on two techniques largely used in image mining. We present clustering methods applied to image analysis, and we introduce the new research direction concerning pattern mining from large collections of images. While considerable advances have been made in image clustering, there is little research dealing with image frequent pattern mining. We will try to understand why.

INTRODUCTION

In recent years, most organisations have been dealing with multimedia data integrating different formats such as images, audio formats, video formats, texts, graphics, or XML documents. For example, a lot of image data have been produced for various professional or domestic domains

such as weather forecasting, surveillance flights, satellites, bio-informatics, biomedical imaging, marketing, tourism, press, Web, and so forth. Such data have been at the disposal of all audiences. Faced with the amount of information produced in numerous domains, there has been a growing demand for tools allowing people to efficiently manage, organise, and retrieve multimedia data.

In this chapter, we focus our attention on the media image. Images may be characterised in terms of three aspects—the volume of the data, the pixel matrix, and the high dimensionality of the data. The first aspect is linked to the huge volume of these data (from a few hundred bytes to several gigabytes for the remote sensing images); the second one reflects the intrinsic nature of the pixel matrix. A pixel or a pixel sequence itself does not mean anything: images do not directly contain any information. Yet the presence of one or more pixel sequences often points to the presence of relevant information. In fact, image interpretation and exploitation need extra relevant information including semantic concepts such as annotations or ontologies, cluster characterisation, and so forth. Today, image and, more generally, multimedia retrieval systems have reached their limits owing to this semantic information absence.

Moreover, in the image retrieval context, a logical indexation process is performed to associate a set of metadata (textual and visual features) with images. These image features are stored in numeric vectors. Their high dimensionality, the third image aspect, constitutes a well known problem. All these different points are, in fact, related to image complexity.

Classical data mining techniques are largely used to analyse alphanumerical data. However, in an image context, databases are very large since they contain strongly heterogeneous data, often not structured and possibly coming from different sources within different theoretical or applicative domains (pixel values, image descriptors, annotations, trainings, expert or interpreted knowledge, etc.). Besides, when objects are described by a large set of features, many of them are correlated, while others are noisy or irrelevant. Furthermore, analysing and mining these multimedia data to derive potentially useful information is not easy. For example, image mining involves the extraction of implicit knowledge, image data relationships, associations between image data, and other data or patterns not explicitly stored in the images.

To circumvent this complexity, we can multiply the number of descriptors. The problem is now to define multidimensional indexes so that searching the nearest neighbours becomes more efficient using the index rather than a sequential search. In the image case, the high dimensionality due to complex descriptors is still an unsolved research problem.

Moreover, another problem is to use external knowledge that could be represented using ontologies or metadata. Taking account of *a priori* knowledge, such as annotation and metadata to build an ontology dedicated to an application, is also a challenge and implies the definition of new descriptors that integrate semantics. As an example, the Web contains many images that are not exploited using the textual part of the Web pages. In this case, the combination of visual and textual information is particularly relevant.

Finally, a crucial task is to organise these large volumes of "raw" data (image, text, etc.) in order to extract relevant information. In fact, decision support systems (DSS) such as data warehousing, data mining, or online analytical processing (OLAP) are evolving to store and analyse these complex data. OLAP and data mining can be seen as two complementary fields. OLAP can easily deal with structuring data before their analysis and with organising structured views. However, this technique is restricted to a simple data navigation and exploration. Data warehouse techniques can help data preprocessing and offer a good structure for an efficient data mining process.

Consequently, new tools must be developed to efficiently retrieve relevant information in