Chapter II
Improving Image Retrieval by Clustering

Dany Gebara
University of Calgary, Canada

Reda Alhajj
University of Calgary, Canada

ABSTRACT

This chapter presents a novel approach for content-fbased image retrieval and demonstrates its applicability on non-texture images. The process starts by extracting a feature vector for each image; wavelets are employed in the process. Then the images (each represented by its feature vector) are classified into groups by employing a density-based clustering approach, namely OPTICS. This highly improves the querying facility by limiting the search space to a single cluster instead of the whole database. The cluster to be searched is determined by applying on the query image the same clustering process OPTICS. This leads to the closest cluster to the query image, and hence, limits the search to the latter cluster without adding the query image to the cluster, except if such request is explicitly specified. The power of this system is demonstrated on non-texture images from the Corel dataset. The achieved results demonstrate that the classification of images is extremely fast and accurate.

INTRODUCTION

Since the early 1990's, there has been considerable research carried out into **content-based image retrieval** (**CBIR**) systems. A few systems have been installed commercially, including Query-By-Image-Content (QBIC) (Niblack, Barber, Equitz, Flickner, Glasman, Petkovic, Yanker, Faloutsos, and Taubin, 1993), the VIR Image Engine (Bach, Fuller, Gupta, Hampapur, Gorowitz, Humphrey, Jain, and Shu, 1996), the AltaVista Photofinder, Multimedia Analysis and Retrieval System (MARS) (Huang, Mehrotra, and Ramchandran, 1996), Photobook (Pentland, Picard, and Sclaroff, 1994), Netra (Ma and Manjunath, 1999), RetrievalWare (Dowe, 1993), etc. Actually, the problem of sorting through images to find a particular object of interest is not new. Whether it is paintings in old museum archives, or browsing through the family albums looking for a particular photograph, extracting information from graphic objects has presented many challenges. With the recent advent and growth of the internet, this problem has been taken to a whole new level. Further, as the hardware needed to capture and store images in digital format has become cheaper and more accessible, the number of people and businesses that have started collecting large numbers of images has grown. The first strategy for dealing with such large collections of images was to tag each image with one or more keywords, allowing existing text-based search systems to work with images. This was a great leap forward, but still had limitations; the biggest of which is that someone had to choose and enter keywords for every image. In addition to being a very tedious task, selection of keywords is a very subjective function. Another method was to sort images by type and place them in file folders much like photographs would be placed in albums. This also suffers from similar drawbacks.

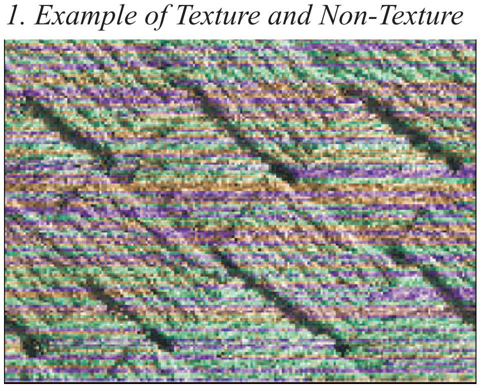
In general, images could be classified into two classes, texture and non-texture. **Texture images** form an important class, where an object within the image is repeated periodically throughout the image. x Some medical images such as X-rays and some topographic images fall under this category. **Non-texture images** tend to have objects of interest clustered in one or more regions of an image. Figure 1 shows one image from each class.

In order to be able to compare images by content, a feature vector (or representative signature) needs to be calculated for each image. This feature vector is the description of the image to the **content-based image retrieval** (**CBIR**) system, which will then conduct its search based on these calculated vectors. Generally, the algorithms used to calculate these feature vectors perform well on some class of images and poorly on others. It therefore follows that a **CBIR** system should classify an image first, and then use an appropriate algorithm based on the classification.

In terms of querying speed, a faster system is naturally preferred. Hence, if there is a way to avoid scanning the entire database every time a query is submitted, this should result in faster responses to the user. **Clustering** can be applied to the calculated feature vectors, where the signatures for similar images are grouped as one cluster. When querying, a **CBIR** system need only to look at a representative for each cluster to narrow the search.

To handle the classification and querying of images better in a more concise and effective way, this chapter proposes a system that combines both wavelet analysis and clustering into the image retrieval

Figure 1. Example of Texture and Non-Texture



Related Content