Computer Vision and Image Understanding 114 (2010) 745–757

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

A content based image retrieval system for a biological specimen collection

Joyita Mallik [a], Ashok Samal [a,*], Scott L. Gardner [b]

[a] Department of Computer Science and Engineering, University of Nebraska-Lincoln, Lincoln, NE 68588-0115, United States
[b] Harold W. Manter Laboratory of Parasitology, University of Nebraska State Museum and School of Biological Sciences, University of Nebraska-Lincoln, Lincoln, NE 68588-0514, United States

ARTICLE INFO

Article history:
Received 1 December 2008
Accepted 11 January 2010
Available online 26 February 2010

Keywords:
CBIR
Biological collections
Shape matching

ABSTRACT

Digital photography and decreasing cost of storing data in digital form has led to an explosion of large digital image repositories. Since the number of images in image databases can be large (millions in some cases) it is important to develop automated tools to search them. In this paper, we present a content based image retrieval system for a database of parasite specimen images. Unlike most content based image retrieval systems, where the database consists of objects that vary widely in shape and size, the objects in our database are fairly uniform. These objects are characterized by flexible body shapes, but with fairly rigid ends. We define such shapes to be FleBoRE (Flexible Body Rigid Extremities) objects, and present a shape model for this class of objects. We have defined similarity functions to compute the degree of likeness between two FleBoRE objects and developed automated methods to extract them from specimen images. The system has been tested with a collection of parasite images from the Harold W. Manter Laboratory for Parasitology. Empirical and expert-based evaluations show that query by shape approach is effective in retrieving specimens of the same class.

1. Introduction

With the advent of digital imaging technology and sharply reduced cost of storage media, there has been an explosion in the number of image databases in recent years [33]. Many of these databases contain millions of images. Browsing through these large image databases to find images of interest is not efficient. Thus, there is a need to develop automated tools to extract useful images from the databases based on the semantic content of the images.

Early techniques for image retrieval from databases were based on searching for text annotations of images, rather than their content directly. However, it is usually difficult to derive a concise and complete textual description of an image. The same image may be interpreted in different ways to produce completely different annotations. The task of annotating the images is not only tedious, but also subjective, sometimes inaccurate and often incomplete. Therefore, text-based image retrieval is neither efficient nor practical for large databases. Content based image retrieval systems use an automated approach to searching and retrieving images from large databases based on image content. The images are searched based on information that can be extracted from them, e.g. color, texture, and shape of the objects.

In this paper, we describe a CBIR system for a collection of images of biological specimen. The objects in an image, i.e. specimen, are characterized by their elongated shapes with a flexible body but rigid extremities. Thus, while the overall shape of multiple instances of the same class varies, there is a uniformity of shape at their two extremities. We label these specimens to be Flexible Body with Rigid Extremities (FleBoRE) objects. In this paper, we describe a model for FleBoRE objects, mechanisms to extract them from specimen images and a framework to retrieve them using a query-by-shape mechanism from a specimen image database. Fig. 1 shows some sample specimen in the image database.

1.1. Motivation

This research is motivated by the need to develop a CBIR system for large collection of parasites in the Harold W. Manter Laboratory (HWML) Parasite Collection. It is one of the top four parasite collections in the world with several million specimens. Researchers from all over the world make use of this collection for research and education. Specimen loans are usually done via mail which is not only expensive, but can result in loss and damage of the specimen. A CBIR system will help bridge the gap between the collections and their users by providing instantaneous access to the specimen. Furthermore, additional tools for image analysis will make the process of specimen-based research more efficient. Search for a specimen by shape or structure will result in images that may point to linkages not otherwise known.

To the best of our knowledge, there is no CBIR system that is focused on databases of parasite specimen characterized by

* Corresponding author. Tel.: +1 402 472 2217; fax: +1 402 472 7767.
E-mail address: samal@cse.unl.edu (A. Samal).

1077-3142/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2010.01.006

Fig. 1. Some sample parasite specimen images in HWML collection.

semi-flexible objects as described here. In particular these objects that have Flexible Bodies and Rigid Extremities (FleBoRE) resemble closely in shape and have only a few specific structural differences. This subtle but important difference is often not modeled in a general CBIR system. For such collection of images, it is therefore necessary to develop models for these objects, devise efficient methods to extract them from images and develop techniques to retrieve those using appropriate querying mechanisms.

2. Related work

In this section we review the work on different aspects of content based image retrieval in various commercial and research systems that is relevant to our work. Most of the domain specific systems that we encountered were related to retrieving biomedical images [2,9,14]. However, the special attributes for medical images, the manner in which they are used are very different from parasite specimen images. For example, in X-ray images of the human spine, the representation of the feature points in the shape of the spine is fundamentally different from the shape representation of the parasite specimen.

In this section, we first discuss some well-known CBIR systems. We then summarize the most common methods of querying in CBIR. Feature selection methods are described next. Finally we discuss methods of computing similarity between objects in two images based on selected features.

2.1. CBIR systems

In this section we briefly discuss some general purpose CBIR systems that have been reported in the literature [41]. Most of them combine multiple features for retrieval of images.

2.1.1. QBIC (query by image content)

This is one of the earliest CBIR systems that was developed by IBM to allow the user to search through image databases using a variety of query mechanisms including query by image and query by sketch [30,41]. Other methods of querying include selection using color palette and texture. First, features related to color, texture and shape of images. A color histogram is typically used to match color features. The texture features include coarseness, contrast and directionality. The shape features include area, circularity, eccentricity, major axis orientation and algebraic moments.

2.1.2. Photobook

Photobook was developed by the MIT Media Laboratory [29,30] and supports both text based and content based image retrieval. Shape and texture features are derived from images. Users query the system by selecting an image from a grid of displayed images. The computed features are then matched to generate the resultant set of images with similar content.

2.1.3. STAR (system for trademark archival and retrieval)

This system uses color and shape features for image retrieval [19]. In this case "shape" is represented as a combination of region-based and outline-based features.

2.1.4. BlobWorld

This is a CBIR system developed at the University of California, Berkeley [6]. The features used in this system include color, texture, location, and shape of regions (blobs). Querying is performed interactively, where a user first selects a region or blob and specifies its importance. The user also indicates the significance of the other features such as color and texture. Retrieval is performed based on the input parameters.

There are a number of other commercial and research CBIR systems in use; however, a full description of all is beyond the scope of the paper. A description and comparative study of different systems can be found in [41].

A number of CBIR systems have been developed for biological specimen in general. Shyu et al. have developed RFLPRetriever, a CBIR system for RFLP (Restriction Fragment Length Polymorphism) images that contain genetic information using biological relevant features in the images [35]. Park and Hwang proposed a portable information bank that can provide a variety of features including content based image retrieval of aquatic plants [27].

Chen et al. describe a CBIR system that is used to study fish specimen images. The shape is represented as landmarks [7]. Experiments were done a specific genus of fish (*Carpiodes*). Murthy et al. also propose a system to identify fish from images [24]. A more general framework for understanding biodiversity using CBIR is proposed by Torres et al. [39].

Several systems to recognize and retrieve images from a database of leaves have been described in [15,25,26,28]. The vein patterns in the leaves and their structure is used as the basis of matching and retrieval. A shape based method to identify leaves is proposed by Shen et al. [34]. Another system to identify plant species from the shape of leaves has been described by Belhumeur et al. [4]. Yet another system based on leave shape to recognize species is proposed by Agarwal et al. [1].

2.2. Querying methods

When searching for images based on textual annotations, retrieval is done by matching the query words to the indexed annotations of the image, using traditional text retrieval methods. Such an approach cannot be used when comparing two images to discover whether they are similar or not. Some common methods of